Sufficient Covariate, Propensity Variable and Doubly Robust Estimation

Hui Guo, Philip Dawid and Giovanni Berzuini

Abstract Statistical causal inference from observational studies often requires adjustment for a possibly multi-dimensional variable, where dimension reduction is crucial. The propensity score, first introduced by Rosenbaum and Rubin, is a popular approach to such reduction. We address causal inference within Dawid's decision-theoretic framework, where it is essential to pay attention to sufficient covariates and their properties. We examine the role of a propensity variable in a normal linear model. We investigate both population-based and sample-based linear regressions, with adjustments for a multivariate covariate and for a propensity variable. In addition, we study the augmented inverse probability weighted estimator, involving a combination of a response model and a propensity model. In a linear regression with homoscedasticity, a propensity variable is proved to provide the same estimated causal effect as multivariate adjustment. An estimated propensity variable may, but need not, yield better precision than the true propensity variable. The augmented inverse probability weighted estimator is doubly robust and can improve precision if the propensity model is correctly specified.

Hui Guo
Centre for Biostatistics, Institute of Population Health, The University of Manchester, Jean McFarlane Building, Oxford Road, Manchester M13 9PL, UK, e-mail: hui.guo@manchester.ac.uk

Philip Dawid
Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK, e-mail: apd25@cam.ac.uk

Giovanni Berzuini
Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy, e-mail: giomanuel_b@hotmail.com

1 Introduction

Causal effects can be identified from well-designed experiments, such as randomised controlled trials (RCT), because treatment assignment is entirely unrelated to subjects' characteristics, both observed and unobserved. Suppose there are two treatment arms in an RCT: treatment group and control group. Then the average causal effect (ACE) can simply be estimated as the outcome difference of the two groups from the observed data. However, randomised experiments, although ideal and to be conducted whenever possible, are not always feasible. For instance, to investigate whether smoking causes lung cancer, we cannot randomly force a group of subjects to take cigarettes. Moreover, it may take years or longer for development of this disease. Instead, a retrospective case-control study may have to be considered. The task of drawing causal conclusion, however, becomes problematic since similarity of subjects from the two groups will rarely hold, e.g., lifestyles of smokers might be different from those of non-smokers. Thus, we are unable to "compare like with like" – the classic problem of confounding in observational studies, which may require adjusting for a suitable set of variables (such as age, sex, health status, diet). Otherwise, the relationship between treatment and response will be distorted, and lead to biased inferences. In general, linear regressions, matching or subclassification are used for adjustment purpose. If there are multiple confounders, especially for matching and subclassification, identifying two individuals with very similar values of all confounders simultaneously would be cumbersome or impossible. Thus, it would be sensible to replace all the confounders by a scalar variable. The propensity score [22] is a popular dimension reduction approach in a variety of research fields.

2 Framework

The aim of statistical causal inference is to understand and estimate a "causal effect", and to identify scientific and in principle testable conditions under which the causal effect can be identified from observational studies. The philosophical nature of "causality" is reflected in the diversity of its statistical formalisations, as exemplified by three frameworks:

1. Rubin's potential response framework [24, 25, 26] (also known as Rubin's causal model) based on counterfactual theory;
2. Pearl's causal framework [16, 17] richly developed from graphical models;
3. Dawid's decision-theoretic framework [6, 7] based on decision theory and probabilistic conditional independence.

In Dawid's framework, causal relations are modelled entirely by conditional probability distributions. We adopt it throughout this chapter to address causal inference; the assumptions required are, at least in principle, testable.

Let X, T and Y denote, respectively, a (typically multivariate) confounder, treatment, and response (or outcome). For simplicity, Y is a scalar and X a multi-

dimensional variable. We assume that T is binary: 1 (treatment arm) and 0 (control arm). Within Dawid's framework, a non-stochastic regime indicator variable F_T, taking values $\emptyset$, 0 and 1, is introduced to denote the treatment assignment mechanism operating. This divides the world into three distinct regimes, as follows:

1. $F_T = \emptyset$: the observational (idle) regime. In this regime, the value of the treatment is passively observed and treatment assignment is determined by Nature.
2. $F_T = 1$: the interventional treatment regime, i.e., treatment T is set to 1 by manipulation.
3. $F_T = 0$: the interventional control regime, i.e., treatment T is set to 0 by manipulation.

For example, in an observational study of custodial sanctions, our interest is in the effect of custodial sanction, as compared to probation (noncustodial sanction), on the probability of re-offence. Then $F_T = \emptyset$ denotes the actual observational regime under which data were collected; $F_T = 1$ is the (hypothetical) interventional regime that always imposes imprisonment; and $F_T = 0$ is the (hypothetical) interventional regime that always imposes probation. Throughout, we assume full compliance and no dropouts, i.e., each individual actually takes whichever treatment they are assigned to. Then we have a joint distribution P_f of all relevant variables in each regime $F_T = f$ ($f = 0, 1, \emptyset$).

In the decision-theoretic framework, causal assumptions are construed as assertions that certain marginal or conditional distributions are common to all regimes. Such assumptions can be formally expressed as properties of conditional independence, where this is extended to allow non-stochastic variables such as F_T [4, 5, 7]. For example, the "ignorable treatment assignment" assumption in Rubin's causal model (RCM) [22] can be expressed as

$$Y \perp\!\!\!\perp F_T | T, \tag{1}$$

read as "Y is independent of F_T given T". However, this condition will be most likely inappropriate in observational studies where randomisation is absent.

Causal effect is defined as the response difference by manipulating treatment, which purely involves interventional regimes. In particular, the population-based average causal effect (ACE) of the treatment is defined as:

$$\text{ACE} := \text{E}(Y|F_T = 1) - \text{E}(Y|F_T = 0), \tag{2}$$

or alternatively,

$$\text{ACE} := \text{E}_1(Y) - \text{E}_0(Y)^1. \tag{3}$$

Without further assumptions, by its definition ACE is not identifiable from the observational regime.

[1] For convenience, the values of the regime indicator F_T are presented as subscripts.

3 Identification of ACE

Suppose the joint distribution of (F_T, T, Y) is known and satisfies (1). Is ACE identifiable from data collected in the observational regime? Note that (1) demonstrates that the distribution of Y given $T = t$ is the same, whether t is observed in the observational regime $F_T = \emptyset$, or in the interventional regime $F_T = t$. As discussed, this assumption would not be satisfied in observational studies, and thus, direct comparison of response from the two treatment groups cannot be interpreted as the causal effect from observational data.

Definition 1. The "face-value average causal effect" (FACE) is defined as:

$$\text{FACE} := \text{E}_\emptyset(Y|T=1) - \text{E}_\emptyset(Y|T=0). \tag{4}$$

It would be hardly true that FACE = ACE, as we would not expect the conditional distribution of Y given $T = t$ is the same in any regime. In fact, identification of ACE from observational studies requires, on one hand, adjusting for confounders, on the other hand, interplay of distributional information between different regimes. One can make no further progress unless some properties are satisfied.

3.1 Strongly sufficient covariate

Rigorous conditions must be investigated so as to identify ACE.

Definition 2. X is a covariate if:

Property 1.

$$X \perp\!\!\!\perp F_T.$$

That is, the distribution of X is the same in any regime, be it observational or interventional. In most cases, X are attributes determined prior to the treatment, for example, blood types and genes.

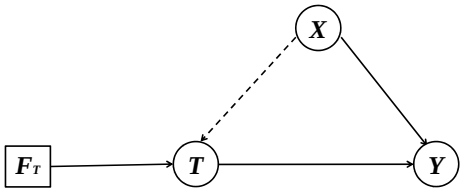
Definition 3. X is a sufficient covariate for the effect of treatment T on response Y if, in addition to Property 1, we have

Property 2.

$$Y \perp\!\!\!\perp F_T | (X, T).$$

Property 2 requires that the distribution of Y, given X and T, is the same in all regimes. It can also be described as "strongly ignorable treatment assignment, given X" [22]. We assume that readers are familiar with the concept and properties of directed acyclic graphs (DAGs). Then Properties 1 and 2 can be represented by means of a DAG as Fig.1. The dashed arrow from X to T indicates that T is partially dependent on X, i.e., the distribution of T depends on X in the observational regime, but not in the interventional regime where $F_T = t$.

Fig. 1 Sufficient covariate



Definition 4. X is a *strongly sufficient covariate* if, in addition to Properties 1 and 2, we have

Property 3. $\mathrm{P}_\emptyset(T = t \mid X) > 0$ with probabilility 1, for $t = 0, 1$.

Property 3 requires that, for any $X = x$, both treatment and control groups are observed in the observational regime.

Lemma 1. *Suppose X is a strongly sufficient covariate. Then, considered as a joint distributions for (Y, X, T), P_t is absolutely continuous with respect to $\mathrm{P}_\emptyset$ (denoted by $\mathrm{P}_t \ll \mathrm{P}_\emptyset$), for $t = 0$ and $t = 1$. That is, for every event A determined by (X, T, Y),*

$$\mathrm{P}_\emptyset(A) = 0 \quad \Longrightarrow \quad \mathrm{P}_t(A) = 0. \tag{5}$$

Equivalently, if an event A occurs with probability 1 under the measure $\mathrm{P}_\emptyset$, then it occurs with probability 1 under the measure P_t ($t = 0, 1$).

Proof. Property 2, expressed equivalently as $(Y, X, T) \perp\!\!\!\perp F_T \mid (X, T)$, asserts that there exists a function $w(X, T)$ such that

$$\mathrm{P}_f(A \mid X, T) = w(X, T)$$

almost surely (a.s.) in each regime $f = 0, 1, \emptyset$. Let $\mathrm{P}_\emptyset(A) = 0$. Then a.s. $[\mathrm{P}_\emptyset]$,

$$0 = \mathrm{P}_\emptyset(A \mid X) = w(X, 1)\mathrm{P}_\emptyset(T = 1 \mid X) + w(X, 0)\mathrm{P}_\emptyset(T = 0 \mid X).$$

By Property 3, for $t = 0, 1$,

$$w(X, t) = 0 \tag{6}$$

a.s. $[\mathrm{P}_\emptyset]$. As $w(X, t)$ is a function of X, it follows that (6) holds a.s. $[\mathrm{P}_t]$ by Property 1. Consequently,

$$w(X, T) = 0 \quad \text{a.s. } [\mathrm{P}_t], \tag{7}$$

since a.s. $[\mathrm{P}_t]$, $T = t$ and $w(X, T) = w(X, t)$ for any bounded function w. Then by (7),

$$\mathrm{P}_t(A) = \mathrm{E}_t\{\mathrm{P}_t(A \mid X, T)\} = \mathrm{E}_t\{w(X, T)\} = 0.$$

Lemma 2. *For any integrable $Z \preceq$ [2] (Y, X, T), and any versions of the conditional expectations,*

$$E_t(Z \mid X) = E_t(Z \mid X, T) \quad \textit{a.s. } [\mathrm{P}_t]. \tag{8}$$

[2] The $\preceq$ symbol is interpreted as "a function of".

Proof. Let $j(X,T)$ be an arbitrary but fixed version of $\mathrm{E}_t(Z \mid X,T)$. Then $j(X,T) = j(X,t)$ a.s. $[\mathrm{P}_t]$, and $j(X,t)$ serves as a version of $\mathrm{E}_t(Z \mid X,T)$ under $[\mathrm{P}_t]$. So

$$\mathrm{E}_t(Z \mid X) = \mathrm{E}_t\{j(X,T) \mid X\} = \mathrm{E}_t\{j(X,t) \mid X\} = j(X,t) \quad \text{a.s. } [\mathrm{P}_t].$$

Thus $j(X,t)$ is a version of $\mathrm{E}_t(Z \mid X)$ under $[\mathrm{P}_t]$ and (8) follows.

Since $\mathrm{E}_t(Z \mid X)$ is a function of X, then by Property 1, $j(X,t)$ is a version of $\mathrm{E}_t(Z \mid X)$ in any regime. Let $g(X,T)$ be some arbitrary but fixed version of $\mathrm{E}_\emptyset(Z \mid X,T)$.

Theorem 1. *Suppose that X is a strongly sufficient covariate. Then for any integrable $Z \preceq (Y,X,T)$, and with notation as above,*

$$j(X,t) = g(X,t) \tag{9}$$

almost surely in any regime.

Proof. By Property 2, there exists a function $h(X,T)$ which is a common version of $\mathrm{E}_f(Z \mid X,T)$ under $[\mathrm{P}_f]$ for $f = 0,1,\emptyset$. Then $h(X,T)$ serves as a version of $\mathrm{E}_\emptyset(Z \mid X,T)$ under $[\mathrm{P}_\emptyset]$, and a version of $\mathrm{E}_t(Z \mid X,T)$ under $[\mathrm{P}_t]$. As $j(X,T)$ is a version of $\mathrm{E}_t(Z \mid X,T)$,

$$j(X,T) = h(X,T) \quad \text{a.s. } [\mathrm{P}_t],$$

and consequently

$$j(X,t) = h(X,t) \quad \text{a.s. } [\mathrm{P}_t].$$

Since $j(X,t)$ and $h(X,t)$ are functions of X, by Property 1

$$j(X,t) = h(X,t) \quad \text{a.s. } [\mathrm{P}_f] \tag{10}$$

for $f = 0,1,\emptyset$. We also have that $g(X,T) = h(X,T)$ a.s. $[\mathrm{P}_\emptyset]$, and so, by Lemma 1, a.s. $[\mathrm{P}_t]$. Then $g(X,t) = h(X,t)$ a.s. $[\mathrm{P}_t]$, where $g(X,t)$ and $h(X,t)$ are both functions of X. By Property 1,

$$g(X,t) = h(X,t) \quad \text{a.s. } [\mathrm{P}_f] \tag{11}$$

for $f = 0,1,\emptyset$. Thus (9) holds by (10) and (11).

3.2 Specific causal effect

Let X be a covariate.

Definition 5. The *specific causal effect* of T on Y, relative to X, is

$$\mathrm{SCE} := \mathrm{E}_1(Y \mid X) - \mathrm{E}_0(Y \mid X).$$

We annotate SCE_X to express SCE as a function of X and write $\mathrm{SCE}(x)$ to indicate that X takes specific value x. Because it is defined in the interventional regimes, SCE has a direct causal interpretation, i.e., $\mathrm{SCE}(x)$ is the average causal effect in the subpopulation with $X = x$.

Although we do not assume the existence of potential responses, when this assumption is made we might proceed as follows. Take X to be the pair $\mathbf{Y} = (Y(1), Y(0))$ of potential responses—which is assumed to satisfy Property 1. Then $\mathrm{E}_t(Y \mid X) = Y(t)$, and consequently

$$\mathrm{SCE}_{\mathbf{Y}} = Y(1) - Y(0),$$

which is the definition of "individual causal effect", ICE, in Rubin's causal model. Thus, although the formalisations of causality are different, SCE in Dawid's decision theoretic framework can be reagarded as a generalisation of ICE in Rubin's causal model.

We can easily prove that, for any covariate X, $\mathrm{ACE} = \mathrm{E}(\mathrm{SCE}_X)$, where the expectation may be taken in any regime. Since by Property 1,

$$\mathrm{E}_\emptyset\{\mathrm{E}_t(Y \mid X)\} = \mathrm{E}_t\{\mathrm{E}_t(Y \mid X)\} = \mathrm{E}_t(Y),$$

for $t = 0, 1$. Thus by subtraction, $\mathrm{ACE} = \mathrm{E}_f(\mathrm{SCE}_X)$ for any regime $f = 0, 1, \emptyset$ and therefore the subscript f can be dropped. Hence, ACE is identifiable from observational data so long as SCE_X is identifiable from observational data. If X is a strongly sufficient covariate, by Theorem 1, $\mathrm{E}_t(Y \mid X)$ is identifiable from the observational regime. It follows that SCE can be estimated from data purely collected in the observational regime. Then ACE expressed as

$$\mathrm{ACE} = \mathrm{E}_\emptyset(\mathrm{SCE}_X) \tag{12}$$

is identifiable, from the observational joint distribution of (X, T, Y). Formula (12) is Pearl's "back-door formula" [17] because by the property of modularity, $\mathrm{P}(X)$ is the same with or without intervention on T and thus can be taken as the distribution of X in the observational regime.

3.3 Dimension reduction of strongly sufficient covariate

Suppose X is a multi-dimensional strongly sufficient covariate. The adjustment process might be simplified if we could replace X by some reduced variable $V \preceq X$, with fewer dimensions—so long as V is itself a strongly sufficient covariate. Now since V is a function of X, Properties 1 and 3 will automatically hold for V. We thus only need to ensure that V satisfies Property 2: that is,

$$Y \perp\!\!\!\perp F_T \mid (V, T). \tag{13}$$

Since two arrows initiate from X in Fig.1, possible reductions may be naturally considered, on the pathways from X to T, and from X to Y. Indeed, the following theorem gives two alternative sufficient conditions for (13) to hold. However, (13) can still hold without these conditions.

Theorem 2. *Suppose X is a strongly sufficient covariate and $V \preceq X$. Then V is a strongly sufficient covariate if either of the following conditions is satisfied:*

(a). **Response-sufficient reduction:**

$$Y \perp\!\!\!\perp X | (V, F_T = t), \tag{14}$$

or

$$Y \perp\!\!\!\perp X | (V, T, F_T = \emptyset), \tag{15}$$

for $t = 0, 1$. It is indicated in (14) that, in each interventional regime, X contributes nothing towards predicting Y once we know V. In other words, as long as V is observed, X need not be observed to make inference on Y. While (15) implies that in the observational regime, knowing X is of no value of predicting Y if V and T are known.

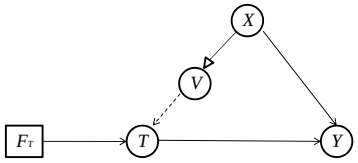
(b). **Treatment-sufficient reduction:**

$$T \perp\!\!\!\perp X | (V, F_T = \emptyset). \tag{16}$$

That is, in the observational regime, treatment does not depend on X conditioning on the information of V.

Proofs of the above reductions were provided in [9]. An alternative proof of (b) can be implemented graphically [9], which results in a DAG as Fig. 2 [3] off which (16) and (13) can be directly read.

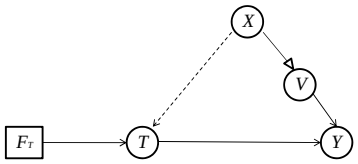
Fig. 2 Treatment sufficient reduction



A graphical approach to (a) does not work since Property 3 is required. However, while not serving as a proof, Fig. 3 conveniently embodies the conditional independencies Properties 1, 2 and the trivial property $V \perp\!\!\!\perp T | (X, F_T)$, as well as (13).

[3] The hollow arrow head, pointing from X to V, is used to emphasise that V is a function of X.

Fig. 3 Response sufficient reduction



4 Propensity analysis

Here we further discuss the treatment-sufficient reduction, which does not involve the response. This brings in the concept of *propensity variable*: a minimal treatment-sufficient covariate, for which we investigate the unbiasedness and precision of the estimator of ACE. Also the asymptotic precision of the estimated ACE, as well as the variation of the estimate from the actual data, will be analysed. In a simple normal linear model that applied for covariate adjustment, two cases are considered: homoscedasticity and heteroscedasticity. A non-parametric approach – subclassification will also be conducted, for different covariance matrices of X of the two treatment arms. The estimated ACE obtained by adjusting for multivariate X and by adjusting for a scalar propensity variable, will then be compared theoretically and through simulations [9].

4.1 Propensity score and propensity variable

The propensity score (PS), first introduced by Rosenbaum and Rubin, is a balancing score [22]. Regarded as a useful tool to reduce bias and increase precision, it is a very popular approach to causal effect estimation. PS matching (or subclassification) method, widely used in various research fields, exploits the property of *conditional (within-stratum) exchangeability*, whereby individuals with the same value of PS (or belonging to a group with similar values of PS) are taken as comparable or exchangeable. We will, however, mainly focus on the application of PS within a linear regression. The definitions of the balancing score and PS given below are borrowed from [22].

Definition 6. A *balancing score* $b(X)$ is a function of X such that, in the observational regime [4], the conditional distribution of X given $b(X)$ is the same for both treatment groups. That is,

$$X \perp\!\!\!\perp T | (b(X), F_T = \emptyset).$$

[4] Rosenbaum and Rubin do not define the balancing score and the PS explicitly for observational studies, although they do aim to apply the PS approach in such studies.

It has been shown that adjusting for a balancing score rather than X results in unbiased estimate of ACE, with the assumption of strongly ignorable treatment assignment [22]. One can trivially choose $b(X) = X$, but it is more constructive to find a balancing score to be a many to one function.

Definition 7. The *propensity score*, denoted by Π, is the probability of being assigned to the treatment group given X in the observational regime:

$$\Pi := \mathrm{P}_{\emptyset}(T = 1 \mid X).$$

We shall use the symbol π to denote a particular realisation of Π. By (16) and Definitions 6 and 7, we assert that PS is the coarsest balancing score. For a subject i, PS is assumed to be positive, i.e., $0 < \pi_i < 1$. Those with the same value of PS are equally likely to be allocated to the treatment group (or equivalently, to the control group), which provides observational studies with the randomised-experiment-like property based on measured X. This is because the characteristics of the two groups with the same or similar PS are "balanced". Therefore, the scalar PS serves as a proxy of multi-dimensional variable X, and thus, it is sufficient to adjust for the former instead of the latter. In observational studies, PS is generally unknown because we do not know exactly which components of X have impact on T and how the treatment is associated with them. However, we can estimate PS from the observational data.

PS analysis for causal inference is based on a sequence of two stages:

Stage 1: PS Estimation. It is estimated by the observed T and X, and normally by a logistic regression of T on X for binary treatment. Note that the response Y is irrelevant at this stage. Because we can estimate PS without observing Y, there is no harm in finding an "optimal" regression model of T on X by repeated trials.

Stage 2: Adjusting for PS. Various adjustment approaches have been developed, e.g., linear regression. If we are unclear about the conditional distribution of Y given T and PS, non-parametric adjustment such as matching or subclassification could be applied instead.

Although two alternatives for dimension reductions have been provided, in practice, this type of reduction may be more convenient in many cases. For example, certain values of the response may occur rarely and only after long observation periods after treatment. In addition, it may sometimes be tricky to determine a "correct" form for a regression model of Y on X, T and F_T. Swapping the positions of X and T, Equation (16) can be re-expressed as

$$X \perp\!\!\!\perp T | (V, F_T = \emptyset), \tag{17}$$

which states that the observational distribution of X given V is the same for both treatment arms. That is to say, V is a *balancing score* for X.

The treatment-sufficient condition (b) can be equivalently interpreted as follows. Consider the family $\mathscr{Q} = \{Q_0, Q_1\}$ consisting of observational distributions of X for the two groups $T = 0$ and $T = 1$. Then Equation (16), re-expressed as (17), says that V is a *sufficient statistic* (in the usual Fisherian sense [8]) for this family. In par-

ticular, a *minimal* treatment-sufficient reduction is obtained as a minimal sufficient statistic for $\mathscr{Q}$: i.e., any variable almost surely equal to a one-one function of the likelihood ratio statistic $\Lambda := q_1(X)/q_0(X)$, where $q_i(\cdot)$ is a version of the density of Q_i.

Definition 8. A *propensity variable* is a minimal treatment-sufficient covariate, or a one-one function of the likelihood ratio statistic Λ.

The concept of a propensity variable is derived from PS which is related to Λ in the following way:

$$\Pi = \mathrm{P}_\emptyset(T = 1 \mid X) = \theta\Lambda/(1-\theta+\theta\Lambda), \tag{18}$$

where $0 < \theta := \mathrm{P}_\emptyset(T = 1) < 1$ by Property 3.

It is entirely possible, from the above discussion, that a different propensity variable will be obtained if we start from a different strongly sufficient covariate.

4.2 Normal linear model (homoscedasticity)

The above theory will be illustrated by a simple example under linear-normal homoscedastic parametric assumptions.

4.2.1 Model construction

Suppose we have a scalar response variable Y, and a $(p \times 1)$ strongly sufficient covariate X that satisfies Properties 1, 2 and 3. Let the conditional distribution of Y given (X, T, F_T) be specified as:

$$Y \mid (X, T, F_T) \sim \mathscr{N}(d + \delta T + b'X, \phi), \tag{19}$$

where the symbol $\sim$ stands for"is distributed as" and the symbol $\mathscr{N}$ stands for normal distribution, with parameters d and δ (scalar), b $(p \times 1)$, and ϕ (scalar). Note that here and in the following models, we assume no interactions between variables in X although interactions can be formally dealt with via dummy variables. Suppose X is a strongly sufficient covariate, then the coefficient δ of T in (19) is the average causal effect ACE, which can be easily proved as follows.

$$\begin{aligned} \mathrm{ACE} &= \mathrm{E}(\mathrm{SCE}_X) = \mathrm{E}\{\mathrm{E}_1(Y \mid X)\} - \mathrm{E}\{\mathrm{E}_0(Y \mid X)\} \\ &= \mathrm{E}(d + \delta + b'X) - \mathrm{E}(d + b'X) = \delta \qquad \text{by} \quad (19). \end{aligned}$$

It is readily seen that the specific causal effect SCE_X is a constant and equals δ.

From (19), the *linear predictor* $\mathrm{LP} := b'X$ satisfies the conditional independence properties in Condition (a) of Theorem 2. Thus, LP is a response-sufficient reduction

of X, and $\mathrm{E}(Y \mid \mathrm{LP}, T) = d + \delta T + \mathrm{LP}$, with coefficient δ of T that does not depend on the regime by virtue of the sufficiency condition.

Now assume that our model for the observational distribution of (T, X) is as follows:

$$\mathrm{P}_\emptyset(T = 1) = \theta \tag{20}$$

$$X \mid (T, F_T = \emptyset) \sim \mathscr{N}(\mu_T, \Sigma) \tag{21}$$

with parameters $\theta \in (0,1)$, $\mu_0(p \times 1)$, μ_1 $(p \times 1)$, and covariance matrix Σ $(p \times p$, positive definite, identical in the two treatment groups). The corresponding marginal distribution of X is a multivariate normal mixture

$$X \mid F_T = \emptyset \sim (1-\theta)\mathscr{N}(\mu_0, \Sigma) + \theta\,\mathscr{N}(\mu_1, \Sigma), \tag{22}$$

in the observational regime, and because we have assumed Property 1 to hold, also in the interventional regime. The observational distribution of T given X is given by (18), with

$$\begin{aligned} \log \Lambda &= \log\{\mathrm{P}_\emptyset(X \mid T = 1)\} - \log\{\mathrm{P}_\emptyset(X \mid T = 0)\} \\ &= -\frac{1}{2}(\mu_1' \Sigma^{-1} \mu_1 - \mu_0' \Sigma^{-1} \mu_0) + \mathrm{LD}, \end{aligned} \tag{23}$$

where

$$\mathrm{LD} := \gamma' X, \tag{24}$$

with

$$\gamma := \Sigma^{-1}(\mu_1 - \mu_0). \tag{25}$$

LD is Fisher's *linear discriminant* [15], best separating the pair of multivariate normal observational distributions for $X \mid T = 0$ and $X \mid T = 1$.

Suppose V is a linear sufficient covariate – a linear function of X that is itself a sufficient covariate. We have proved that the coefficient of T in the observational linear regression of Y on T and V is δ [9]. From (23) we see that LD is a propensity variable which is a linear strongly sufficient covariate. We deduce that under the given distributions, the coefficient of T in the observational regression of Y on T and LD is δ.

Theorem 3. *The coefficient of T in the linear regression of Y on (T, LD) is the same as that in the linear regression of Y on (T, X).*

Theorem 3 states that it is algebraically true that X and Fisher's linear discriminant LD generate identical coefficient of T in linear regressions, which does not have a direct link to the regimes and causality whatsoever. In our linear normal model, δ is interpreted as ACE and can be identified from the observational data simply because we have assumed that X is a strongly sufficient covariate. Applying Theorem 3 to the empirical distribution of (Y, T, X) from a sample, we deduce Corollary 1 as follows.

Corollary 1. *Suppose we have data on (Y, T, X) for a sample of individuals. Let LD^* be the sample linear discriminant for T based on X. Then the coefficient of T*

in the sample linear regression of Y on T and LD^* *is the same as that in the sample linear regression of Y on T and X.*

Rosenbaum and Rubin [22] (§ 3.4) also give this result with a brief non-causal argument: whenever the sample dispersion matrix is used in both the form of LD and regression adjustment, the estimated coefficient of T must be the same.

As discussed [9], here is a paradox: we regard adjustment for the propensity variable as an adjustment for the treatment assignment process, by regressing Y on T and the estimated propensity variable LD^*. However, from the result of Corollary 1, it appears that what we actually adjust for is the full set of covariates X, which makes the treatment assignment process completely irrelevant.

4.2.2 Precision in propensity analysis

One might intuitively think that the precision of the estimated ACE would be improved if we were to adjust for a scalar variable — the sample-based propensity variable LD^*, rather than p-dimensional variable X. However, Corollary 1 tells us that adjusting for LD^* does not increase the precision of our estimator. In fact, whether one adjusts for LD^* and for all the p predictors makes *absolutely no difference* to our estimate, and thus, to its precision. Similar conclusions have been drawn in [10, 33, 31]. Our intuition is that the increased precision obtained by regressing on V is offset by the overfitting error involved in selecting V.

Previous evidence [21, 11, 28] supports the claim that the estimated propensity variable outperforms the true propensity variable. That is, adjusting for the former yields higher precision of the estimated ACE than the latter. These two types of adjustment correspond to regressing Y on (T, LD) and on (T, LD^*) in our model and both provide an unbiased estimator of ACE. The claim obviously cannot be always valid by simply considering a special case: $\mathrm{LD} = \mathrm{LP}$, because by Corollary 1, regressing on LD^* is the same as adjusting for LP^*, which by the Gauss-Markov theorem will be less precise than regressing on the true linear predictor LP (or equivalently LD). Nevertheless, the claim is likely to hold when LD is not highly correlated with LP because LD is a less precise response predictor.

4.2.3 Asymptotic variance analysis

To gain a closer insight into the variance of the estimated ACE, by adjusting for the true propensity variable PV (if known) and the estimated propensity variable EPV, we consider a toy example in which the parameters in (19), (20) and (21) are set as follows:

$$p = 2, \quad P_\emptyset(T = 1) = \theta \in (0,1), \quad b = (b_1, b_2)',$$

the covariance matrix Σ is diagonal with identical entries τ, and

$$\mathrm{E}(X_2 \mid T = 1) = \mathrm{E}(X_2 \mid T = 0) = \mathrm{E}(X_2) \tag{26}$$

By the setting of Σ, we see that

$$X_1 \perp\!\!\!\perp X_2 \mid T. \tag{27}$$

It is also clear that the true PV is just X_1, by minimal treatment-sufficient reduction and related equations (23), (24), (25). The conditions according to our model setting are expressed by a DAG as Fig. 4.

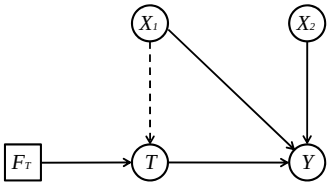


Fig. 4 Propensity variable X_1 and response predictor $X = (X_1, X_2)$

In practice, all the parameters are unknown, and consequently the exact form of PV is not known. What one would normally do is adjust for the whole set of the observed X, which is equivalent to adjusting for LD* (or EPV) in the linear regression approach by Corollary 1. In particular, two linear regressions are considered as follows:

$\underline{M_0}$: Y on (T, X),

$\underline{M_1}$: Y on (T, X_1).

Then the design matrix is $(1, T, X_1, X_2)'$ for M_0 and $(1, T, X_1)'$ for M_1. Let $\widehat{\beta_{M_0}}$ and $\widehat{\beta_{M_1}}$, respectively, be the least square estimators of the parameters in M_0 and M_1. The asymptotic variance of $\widehat{\beta_{M_0}}$ for sample size n is then given as:

$$\mathrm{Var}_{.asy}(\widehat{\beta_{M_0}}) = \frac{A^{-1}\mathrm{Var}(Y \mid T, X)}{n} = \frac{A^{-1}\phi}{n},$$

where

$$A = \begin{pmatrix} 1 & \theta & \mathrm{E}(X_1) & \mathrm{E}(X_2) \\ \theta & \theta & \mathrm{E}(TX_1) & \mathrm{E}(TX_2) \\ \mathrm{E}(X_1) & \mathrm{E}(TX_1) & \mathrm{E}({X_1}^2) & \mathrm{E}(X_1X_2) \\ \mathrm{E}(X_2) & \mathrm{E}(TX_2) & \mathrm{E}(X_1X_2) & \mathrm{E}({X_2}^2) \end{pmatrix}.$$

By solving A^{-1} and extract the (2, 2)th element which is variance multiplier of the coefficient of T, we have that

$$\mathrm{Var}_{.asy}(\widehat{\delta_{M_0}}) = \frac{(W_{X_1X_1}W_{X_2X_2} - {W_{X_1X_2}}^2)\phi}{n\theta(1-\theta)(V_{X_1X_1}V_{X_2X_2} - {V_{X_1X_2}}^2)},$$

where

$$W_{X_1X_2} = \mathrm{E}(X_1X_2) - \mathrm{E}(X_1)\mathrm{E}(X_2) = \mathrm{Cov}(X_1, X_2),$$

$$V_{X_1X_2} = \mathrm{E}(X_1X_2) - \theta\mathrm{E}(X_1 \mid T=1)\mathrm{E}(X_2 \mid T=1) - (1-\theta)\mathrm{E}(X_1 \mid T=0)\mathrm{E}(X_2 \mid T=0),$$

$$W_{X_1X_1} = \mathrm{E}({X_1}^2) - [\mathrm{E}(X_1)]^2 = \mathrm{Var}(X_1),$$

$$W_{X_2X_2} = \mathrm{E}({X_2}^2) - [\mathrm{E}(X_2)]^2 = \mathrm{Var}(X_2),$$

and

$$V_{X_1X_1} = \mathrm{E}({X_1}^2) - \theta[\mathrm{E}(X_1 \mid T=1)]^2 - (1-\theta)[\mathrm{E}(X_1 \mid T=0)]^2,$$

$$V_{X_2X_2} = \mathrm{E}({X_2}^2) - \theta[\mathrm{E}(X_2 \mid T=1)]^2 - (1-\theta)[\mathrm{E}(X_2 \mid T=0)]^2.$$

By (26),

$$V_{X_2X_2} = \mathrm{Var}(X_2) = W_{X_2X_2}$$

and

$$V_{X_1X_2} = \mathrm{Cov}(X_1, X_2) = W_{X_1X_2},$$

where, by (27),

$$\mathrm{Cov}(X_1, X_2) = \mathrm{E}\{\mathrm{Cov}(X_1 \mid T, X_2 \mid T)\} + \mathrm{Cov}\{\mathrm{E}(X_1 \mid T), \mathrm{E}(X_2 \mid T)\} = 0.$$

Hence,

$$\mathrm{Var}_{.asy}(\widehat{\delta_{M_0}}) = \frac{\phi\mathrm{Var}(X_1)/[n\theta(1-\theta)]}{\mathrm{E}({X_1}^2) - \theta[\mathrm{E}(X_1 \mid T=1)]^2 - (1-\theta)[\mathrm{E}(X_1 \mid T=0)]^2}. \quad (28)$$

For M_1, by (27),

$$\begin{aligned}\mathrm{Var}_{.asy}(\widehat{\delta_{M_1}}) &= \frac{W_{X_1X_1}}{n\theta(1-\theta)V_{X_1X_1}} \cdot \mathrm{Var}(Y \mid T, X_1) \\ &= \frac{W_{X_1X_1}}{n\theta(1-\theta)V_{X_1X_1}} \cdot \{\phi + {b_2}^2\mathrm{Var}(X_2 \mid T, X_1)\} \\ &= \frac{(\phi + {b_2}^2\tau)\mathrm{Var}(X_1)/[n\theta(1-\theta)]}{\mathrm{E}({X_1}^2) - \theta[\mathrm{E}(X_1 \mid T=1)]^2 - (1-\theta)[\mathrm{E}(X_1 \mid T=0)]^2}. \quad (29)\end{aligned}$$

Comparing (28) and (29), we have that $\mathrm{Var}_{.asy}(\widehat{\delta_{M_0}}) < \mathrm{Var}_{.asy}(\widehat{\delta_{M_1}})$ unless X_2 is random noise rather than the linear predictor i.e., $b_2 = 0$ which equalises the two asymptotic variances.

Lemma 3. *Under the given distributional assumptions (19), (20) and (21), suppose the propensity variable* LD *is not the same as the linear predictor* LP*, and* LD *is independent of variables that are merely response predictors. Then the asymptotic variance of the estimated ACE from the linear regression by adjusting for the estimated propensity variable* LD^* *is more precise than that by adjusting for the population propensity variable* LD*.*

4.2.4 Simulations

Simulations are carried out for numerical illustration. Suppose we have the following true values for the paramters in (19), (20) and (21): $p = 2, d = 0, \delta = 0.5, b = (0,1)', \phi = 1, \theta = 0.5, \mu_1 = (1,0)', \mu_0 = (0,0)'$, $\Sigma = I_2$.

Then the population linear predictor is $\mathrm{LP} = X_2$, with

$$Y \mid (X, T, F_T) \sim \mathscr{N}(\frac{1}{2}T + X_2, 1),$$

while the population linear discriminant $\mathrm{LD} = X_1$ which is not predictive to Y. Since for any regime $f = 0, 1, \emptyset$,

$$\mathrm{E}_f(Y \mid X_1, T) = \mathrm{E}_f\{\mathrm{E}_f(Y \mid X, T) \mid X_1, T\} = \frac{1}{2}T$$

and

$$\mathrm{Var}_f(Y \mid X_1, T) = \mathrm{E}_f\{\mathrm{Var}_f(Y \mid X, T) \mid X_1, T\} + \mathrm{Var}\{\mathrm{E}_f(Y \mid X, T) \mid X_1, T\} = 2.$$

The conditional distribution of Y given (X_1, T), for any regime, is then given by

$$Y \mid (X_1, T, F_T) \sim \mathscr{N}(\frac{1}{2}T, 2).$$

To investigate the performance of the population-based as well as sample-based LP and PV, we now consider four linear regression models:

M_0: Y on T and X ($X = (X_1, X_2)$),
M_1: Y on T and X_1,
M_2: Y on T and X_2,
M_3: Y on T and LD^*,

where M_0 is the full model with all parameters unknown. In M_1, by setting $b_2 = 0$, the true linear discriminant $\mathrm{LD} = X_1$ is fitted. While fitting the true linear predictor $\mathrm{LP} = X_2$, equivalent to setting $b_1 = 0$, we get M_2. Note that all these models are "true". For M_1 the true value of b_1 is 0, and the true residual variance is 2, as against 1 for M_0 and M_2. Finally, for any dataset with no information of parameters, we construct the estimated propensity variable LD^*, and then fit the model M_3.

In each model M_k, for $k = 0, 1, 2, 3$, the least-squares estimator $\hat{\delta}_k$ is unbiased for $\delta = 0.5$. By the Gauss-Markov theorem and Corollary 1,

$$\mathrm{Var}(\hat{\delta}_0) = \mathrm{Var}(\hat{\delta}_3) \geq \mathrm{Var}(\hat{\delta}_2).$$

Asymptotically, we have that $\mathrm{Var}.asy(\hat{\delta}_0) = \mathrm{Var}.asy(\hat{\delta}_3) = 5/n$, $\mathrm{Var}.asy(\hat{\delta}_2) = 4/n$, and $\mathrm{Var}.asy(\hat{\delta}_1) = 10/n$. It is indeed asymptotically less precise to adjust for PV than for its estimate in our model, which is in accordance with Lemma 3.

For the sample analysis, 200 simulated datasets are generated, each of size $n = 20$. Shown in Fig. 5 are the empirical distributions of $\hat{\delta}_k$ for all four models. Unsurprisingly, in terms of precision (from high to low), first comes the LP; next

is the estimated propensity variable LD^* (or the estimated linear predictor LP^*), or equivalently, X $(= (X_1, X_2))$; and last comes the true propensity variable $\mathrm{LD} = X_1$.

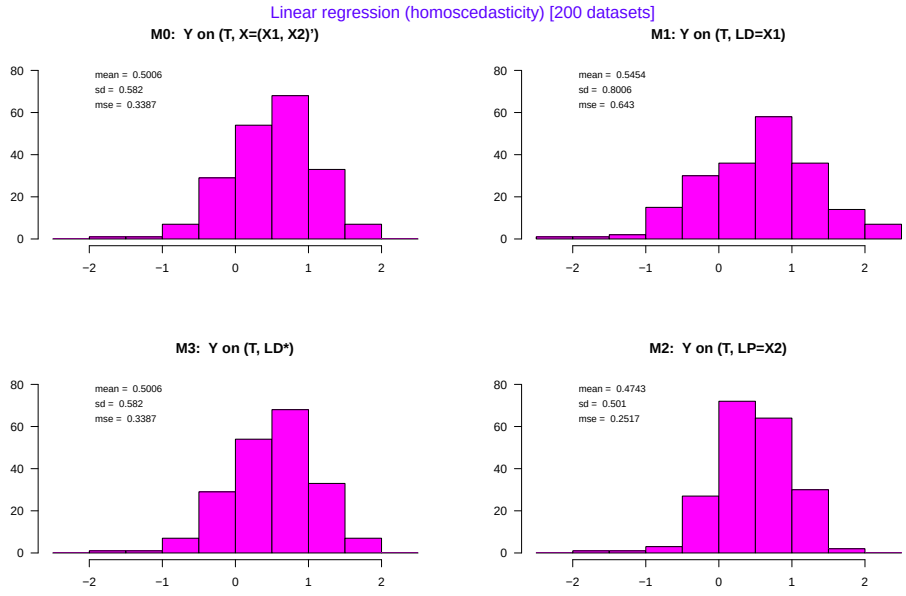


Fig. 5 Estimates of ACE by regression on (clockwise): 1. X_1 and X_2. 2. population linear discriminant (propensity variable) X_1. 3. population linear predictor X_2. 4. estimated linear discriminant (propensity variable) LD^* .

4.3 Normal linear model (heteroscedasticity)

Investigation in the homoscedasticity case is simple because PV is equivalent to LD, where linearity makes analysis straightforward. If covariance matrices of the conditional distribution of X for the two treatment groups are not identical, it turns out that adjusting for PV is not appropriate.

Suppose now that, keeping all other distributional assumptions of §4.2 unchanged, (21) is re-specified as:

$$X \mid (T, F_T = \emptyset) \sim \mathscr{N}(\mu_T, \Sigma_T)$$

with different covariance matrices Σ_0 and Σ_1 for $T=0$ and $T=1$. The distribution of X in all regimes then becomes

$$X \mid F_T \sim (1-\theta)\mathcal{N}(\mu_0,\Sigma_0)+\theta\,\mathcal{N}(\mu_1,\Sigma_1).$$

Accordingly,

$$\log\Lambda = c + \mathrm{QD}$$

where

$$c = -\frac{1}{2}\{\log(\det\Sigma_1)-\log(\det\Sigma_0)+\mu_1'\Sigma_1^{-1}\mu_1-\mu_0'\Sigma_0^{-1}\mu_0\}$$

and

$$\mathrm{QD} := \left(\Sigma_1^{-1}\mu_1-\Sigma_0^{-1}\mu_0\right)'X-\frac{1}{2}X'\left(\Sigma_1^{-1}-\Sigma_0^{-1}\right)X. \tag{30}$$

QD is the *quadratic discriminant* including a linear term and a quadratic term of X, distinguishing the observational distributions of X given $T=0,1$. We see that QD is a minimal treatment-sufficient covariate, and thus a PV but no longer a linear function of X.

Because of the balancing property of PS (or PV), it now follows that ACE $=$ $\mathrm{E}(\mathrm{SCE}_{\mathrm{QD}})$, with

$$\mathrm{SCE}_{\mathrm{QD}} = \mathrm{E}_1(Y \mid \mathrm{QD}) - \mathrm{E}_0(Y \mid \mathrm{QD}).$$

Since QD is quadratic in X, Y is no longer linear in QD, the coefficient of T by adjusting for PV (= QD) in the linear regression does not provide exact ACE. However, as computation of the expectations in the above formula is non-trivial, one might wish to replace $\mathrm{E}_{\emptyset}(Y \mid T,\mathrm{QD})$ by the linear regression of Y on (T,QD), and approximate the estimated ACE. Alternatively, one can take non-parametric approaches such as matching or subclassification on QD [23]. A number of papers on various matching approaches for causal effects have been collected in [27]. More recently, statistical software becomes available for multivariate and PS matching in R [30].

Now we discuss subclassifications and linear regressions based on QD, compared to linear regressions based on LP and LD. The linear discriminant is again in the form

$$\mathrm{LD} = (\mu_1-\mu_0)'\Sigma^{-1}X,$$

but with $\Sigma=(1-\theta)\Sigma_0+\theta\Sigma_1$, the sum of the weighted dispersion matrices of the two treatment groups. Form the formulae of QD and LD, we conclude that it is LD that comprises all variables with expectations depending on T. In a DAG representation of this scenario, each of such variables must have an arrow pointing to T. However, the genuine PV (= QD) may depend on all the components of X, according to its quadratic term in (30). Only with homoscedasiticity, PV is equivalent to LD and includes all variables associated with T.

Although LD is not a sufficient covariate here, Theorem 3 still applies. It enables us to identify ACE from the linear regression of Y on (T,LD), which is equivalent to the linear regression of Y on (T,X). However, other authors claim that only if LD is highly correlated with PS, adjustment for LD works well in regressions [22].

This may attribute to different scenarios considered, i.e., in our model Y is linearly related to X while non-linear in X in theirs.

4.3.1 Simulations

Simulated data is based on the above model, with the parameters: $p = 20$, $d = 0$, $\delta = 0.5$, $\theta = 0.5$, $b = (0,1,\ldots,0)'$, $\mu_0 = (0,\ldots,0)'$, and $\mu_1 = (0.5,0,0,\ldots,0)'$. Also, Σ_0 is set, diagonally, to 0.8 for the first ten entries and to 1.3 for the remaining entries, and Σ_1 the identity matrix.

We then have, for the population, that

$$\mathrm{LD} = \frac{5}{9}X_1,$$

$$\mathrm{PV} = \mathrm{QD} = \frac{1}{2}X_1 + \frac{1}{8}\sum_{i=1}^{10} X_i^2 - \frac{3}{26}\sum_{j=11}^{20} X_j^2,$$

and $\mathrm{LP} = X_2$. By estimating μ_0 and μ_1, Σ_0 and Σ_1 from observed data, we can compute sample-based LD^* and QD^*.

The results from 200 simulated datasets, each of size 500, are given in Fig. 6. The first three plots (clockwise) are from the linear regressions of Y on, respectively (T, X_2), (T, LD), and (T, QD). The last plot is the result of subclassification on PV (= QD). That is, 500 observations are divided into 5 subclasses with equal number of observations in each, based on the values of QD. Within each subclass, units from the two treatment groups are roughly comparable such that the average difference of the response may be interpreted as the estimated SCE. Then ACE is estimated by summing over SCEs, each weighted by $1/5$. Note that the sample size has increased, since we must have at least one observation for each treatment in each subclass.

Since LD and QD are practically unknown, they need be estimated from the observed data. Also, we do not know exactly the response predictors or the confounders, full set of the observed X may have to be used for analysis.

Fig. 7 gives the results from the same 200 datasets as above. Again, the first three plots are the results of linear regressions of Y, but on, respectively, (T, X), (T, LD^*), and (T, QD^*), where LD^* and QD^* are the sample linear and quadratic discriminants. Shown in the last plot is the result of subclassification on EPV (= QD^*). Unsurprisingly, by comparing the mean, standard deviation and mean squared error of the estimated ACE, regression of Y on $(T, \mathrm{LP} = X_2)$ comes the best among all eight approaches in Fig. 6 and Fig. 7. Regressing on (T, X) is no better than regressing on (T, X_2) because all variables except X_2 in X are not predictors but noise of Y. In confirmation of the theory in §4.2.1, regressing on LD^*, rather than on X, has absolutely no effect on the estimated ACE. LD^* outperforms LD because the latter does not contain the response predictor. Regressions on LD, QD, and on QD^* are roughly equal, because apart from X_1, the distributions of the remaining 19 variables are identical, with rather small multipliers. Thus, the two quadratic terms in QD are roughly the same, and $\mathrm{QD} \approx \frac{1}{2}X_1$ works approximately as a function of

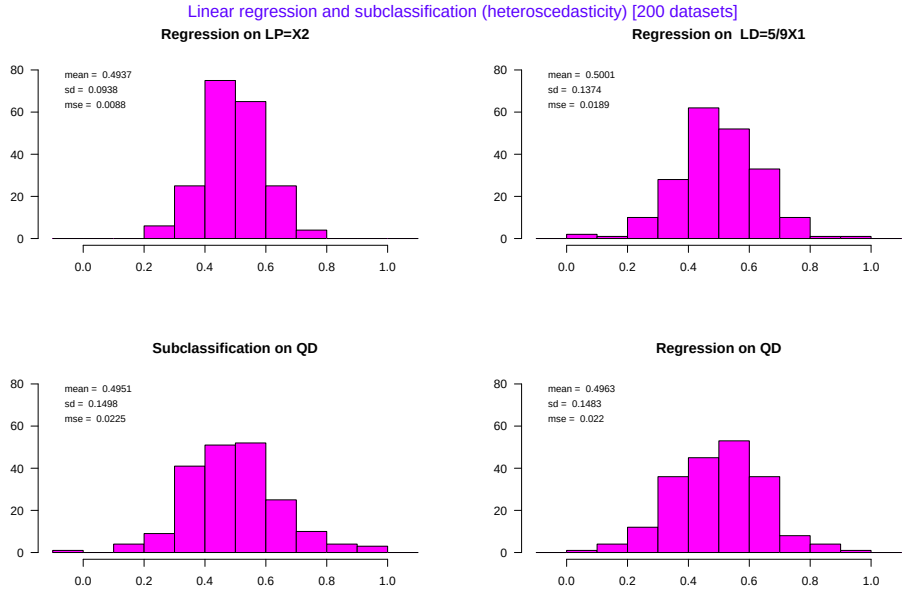


Fig. 6 Estimates of ACE by 4 different methods (clockwise): 1. Regression on population linear predictor $\mathrm{LP} = X_2$. 2. Regression on population linear discriminant $\mathrm{LD} = \frac{5}{9}X_1$. 3. Regression on population quadratic discriminant (propensity variable) QD. 4. Subclassification on QD.

a single variable X_1. Last comes subclassification on the quadratic PV, particularly when it is estimated.

4.4 Propensity analysis in logistic regression

As already investigated, propensity analysis in linear regression is fairly straightforward. In many cases, however, response Y is not linear in X. We know that despite its name, generalised linear model (GLM) is not a linear model, because it is a non-linear function of the response that is linearly related to its predictors. Logistic regression is widely applied as a type of GLM if the response is binary. For example, doctors often record the outcome of a surgery on a patient as either “cured” or “not cured”. Next, a logistic model is used in our illustrative study.

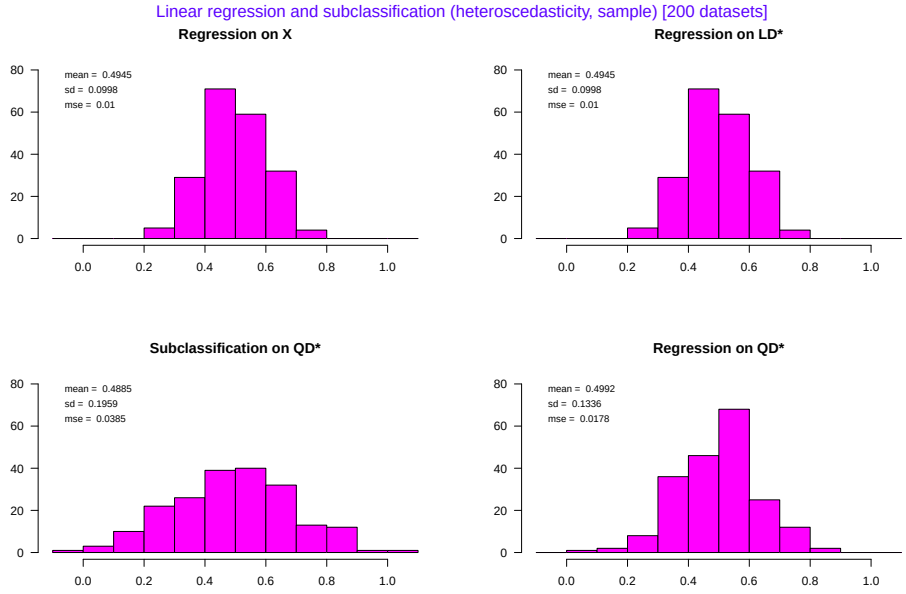


Fig. 7 Estimates of ACE by 4 different methods (clockwise): 1. Regression on sufficient covariate X. 2. Regression on sample linear discriminant LD*. 3. Regression on sample quadratic discriminant (propensity variable) QD*. 4. Subclassification on QD*.

4.4.1 Model construction

For simplicity, suppose that $Y, T\,(1\times 1)$ and $X\,(p\times 1)$ are all binary and components of X are mutually independent, The joint distribution of (F_T, X, T, Y) is constructed as follows:

$$X \mid F_T \sim Ber(\pi) \tag{31}$$

$$\text{logit}\{\mathrm{P}_\emptyset(T \mid X)\} = c + a'X \tag{32}$$

$$\text{logit}\{\mathrm{P}_f(Y \mid T, X)\} = d + \delta T + b'X, \tag{33}$$

for $f = 0, 1, \emptyset$; and π is $(p\times 1)$. Property 3 and $P_f(Y=1 \mid T, X) \in (0,1)$ are required such that (32) and (33) are well-defined.

It is immediately seen that X is a strongly sufficient covariate and

$$\begin{aligned}\text{ACE} &= \mathrm{E}_\emptyset\{\mathrm{E}_1(Y \mid X)\} - \mathrm{E}_\emptyset\{\mathrm{E}_0(Y \mid X)\} \\ &= \mathrm{E}_\emptyset\{\mathrm{P}_\emptyset(Y \mid T=1, X)\} - \mathrm{E}_\emptyset\{\mathrm{P}_\emptyset(Y \mid T=0, X)\}\end{aligned}$$

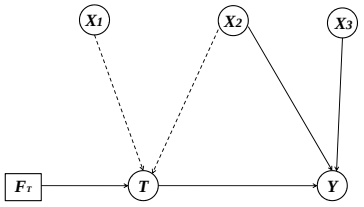
$$= \mathrm{E}_{\emptyset}\{\frac{1}{1+e^{-(d+\delta+b'X)}} - \frac{1}{1+e^{-(d+b'X)}}\}.$$

If the parameters are set as follows:

$$p=3, \quad \pi=(\pi_1,\pi_2,\pi_3)', \quad a=(a_1,a_2,0)', \quad b=(0,b_2,b_3)', \tag{34}$$

then the response predictor is $b_2X_2+b_3X_3$ and PV $= a_1X_1+a_2X_2$. The conditional independence properties in our model can be read off Fig. 8. Then we have that

Fig. 8 DAG for the logistic model



$$\mathrm{logit}\{P(Y \mid T,X_2,X_3)\} = \mathrm{logit}\{P(Y \mid T,X)\} = d+\delta T+b_2X_2+b_3X_3,$$

and

$$\begin{aligned} P(Y=1 \mid T,X_1,X_2) &= P(Y=1 \mid T,X_2) \\ &= \mathrm{E}\{\frac{1}{1+e^{-(d+\delta T+b_2X_2+b_3X_3)}} \mid T,X_2\} \\ &= \frac{\pi_3}{1+e^{-(d+\delta T+b_2X_2+b_3)}} + \frac{1-\pi_3}{1+e^{-(d+\delta T+b_2X_2)}} \end{aligned}$$

which does not depend on X_1. And we have that

$$\begin{aligned} \mathrm{ACE} \;=\; & \pi_2\pi_3\{\frac{1}{1+e^{-(d+\delta+b_2+b_3)}} - \frac{1}{1+e^{-(d+b_2+b_3)}}\} \\ & + (1-\pi_2)\pi_3\{\frac{1}{1+e^{-(d+\delta+b_3)}} - \frac{1}{1+e^{-(d+b_3)}}\} \\ & + \pi_2(1-\pi_3)\{\frac{1}{1+e^{-(d+\delta+b_2)}} - \frac{1}{1+e^{-(d+b_2)}}\} \\ & + (1-\pi_2)(1-\pi_3)\{\frac{1}{1+e^{-(d+\delta)}} - \frac{1}{1+e^{-d}}\}, \end{aligned} \tag{35}$$

which is determined by d,δ,b_2,b_3, π_2 and π_3. This extremely simple example, with only three components of X that are all binary, already results in a complicated form

for ACE, which would be even worse for high dimensional X and various types of variables. Next, instead of simulation, we conduct propensity analysis on real data.

4.4.2 Propensity analysis of custodial sanctions study

We illustrate the method with the aid of a study involving 511 subjects sentenced to prison in 1980 by the California Superior Court, and 511 offenders sentenced to probation following conviction for certain felonies [2]. These probationers were matched to the prisoners on county of conviction, condition offence type and risk of imprisonment quantitative index, so as to bring into the final sample the most serious offenders on probation and the least serious offenders sentenced to prison. The structure of this study corresponds to the (partially matched) case-control design. In fact, this is analogous to the regression discontinuity designs where only observations near the cut-off of the risk score are included for causal effect analysis [13]. We were to compare the average causal effect of judicial sanction (probation or prison) on the probability of re-offence. We specify variables as follows.

- Treatment T: taking values 0 (probation) and 1 (prison);
- Response Y: occurrence of recidivism (re-offence);
- Pre-treatment variable X: including 17 carefully selected non-collinear variables that we can reasonably assume to make X a strongly sufficient covariate.

Simple random multiple imputation by bootstrapping (R package: mi) was applied to deal with missing data. We then considered two logistic regressions for the imputed data:

1. Y on (T,X), where X includes all the 17 variables.
2. Y on $(T,$ EPS), where EPS is the propensity score estimated from the logistic regression of T on all the 17 variables. In selecting these variables, we took advantage of the possibility of trying various sets of covariates in the model, without inflating the type I error since these regressions do not involve the response information. The distribution densities of the two treatment groups are shown in Fig. 9, where we see a large overlapping area.

Shown in Tab. 1 are the results. In this case, regression on the full set of X and on the estimated PS makes little difference, since the summary statistics from the two approaches are quite similar. Although the negative values of both the coefficients imply reduced re-offence for the imprisonment, they are not statistically significant.

Table 1 Coefficients of judicial sanction ("prison" with respect to "probation") from logistic regressions: 1. Y on (T,X); 2. Y on $(T,$ EPS)

Regression	Coefficient	Standard Error	p-value
Y on (T,X)	-0.1631	0.1579	0.3014
Y on $(T,$ EPS)	-0.1713	0.1503	0.2545

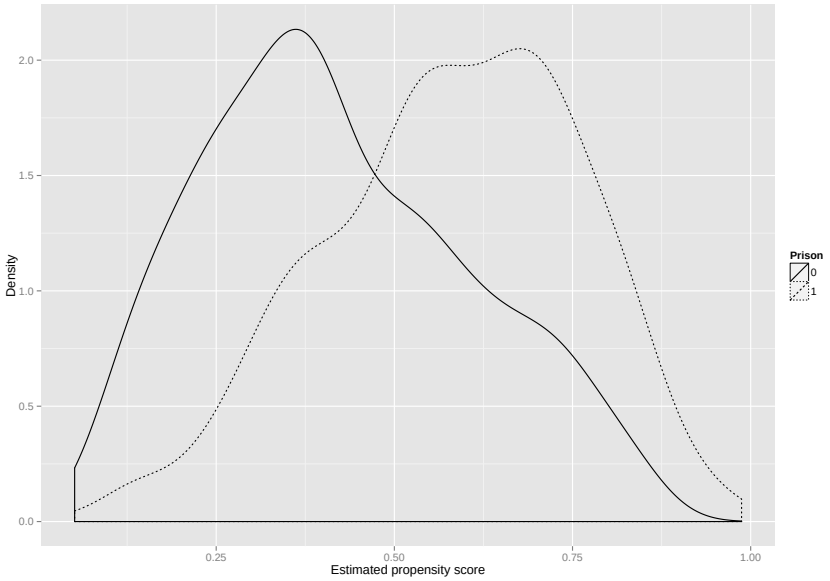


Fig. 9 Distribution density comparison of the estimated propensity score: prison vs. probation.

5 Double robustness

Since the underlying response regression model (RRM): $Y \mid (X, T, F_T = \emptyset)$ and the propensity model (PM): $T \mid (X, F_T = \emptyset)$ are most likely unknown, one may specify parametric models based on previous experience. Moreover, as discussed in § 3.3, a strongly sufficient covariate can be reduced by two alternative approaches from specified models, which enables estimating ACE by either method as follows:

1. Adjustment for response predictors from correctly specified RRM;
2. Adjustment for a PV (or PS) from correctly specified PM, either in response regression (if RRM is correctly specified), or otherwise, by non-parametric approaches, e.g., matching.

Due to lack of knowledge, it may well be that *at least one* model is misspecified. Little could be done if both models are wrong. Thus, our interest is to find a single estimator that produces a good estimate, given that at least one model is correct.

ACE is normally estimated from the observed data. Suppose there are n individuals in an observational study. Observations (x_i, t_i, y_i), where $i = 1, .., n$, are generated from the joint distribution (X_i, T_i, Y_i) that are independent and identically distributed. The estimation of the ACE requires estimates of the expected response for both treatment groups assigned by intervention. We have already demonstrated that, within the DT framework, ACE is identifiable from pure observational data if X is a

strongly sufficient covariate. Here, X is again assumed to be strongly sufficient and thus satisfies Properties 1, 2 and 3.

5.1 Augmented inverse probability weighted estimator

To construct the augmented inverse probability weighted (AIPW) estimator, we discuss two scenarios:

- **Correct RRM**: Suppose that we know the RRM. For convenience, we write $\mathrm{E}_t(Y)$ as μ_t, so
$$\mu_t = \mathrm{E}_\emptyset[\mathrm{E}_\emptyset(Y \mid X, T = t)] \tag{36}$$
since X is strongly sufficient. Hence, in observational studies, $\mathrm{E}_\emptyset(Y \mid X, T = t)$ is an unbiased estimator of μ_t, for $t = 0, 1$. Consequently, $\mathrm{E}_\emptyset(Y \mid X, T = 1) - \mathrm{E}_\emptyset(Y \mid X, T = 0)$ is an unbiased estimator of ACE.
- **Correct PM**: Consider that the PM is correct, i.e., $\pi(X) = \mathrm{P}_\emptyset(T = 1 \mid X)$.

Lemma 4. *Suppose that the propensity model is correct and that X is a strongly sufficient covariate. Then*
$$\mathrm{ACE} = \mathrm{E}_\emptyset\{\frac{T}{\pi(X)}Y\} - \mathrm{E}_\emptyset\{\frac{1-T}{1-\pi(X)}Y\}, \tag{37}$$
where $\mathrm{E}_\emptyset\{\frac{T}{\pi(X)}Y\} = \mu_1$ and $\mathrm{E}_\emptyset\{\frac{1-T}{1-\pi(X)}Y\} = \mu_0$.

Proof.
$$\begin{aligned}
\mathrm{E}_\emptyset\{\frac{T}{\pi(X)}Y\} &= \mathrm{E}_\emptyset\{\mathrm{E}_\emptyset(\frac{T}{\pi(X)}Y \mid X)\} = \mathrm{E}_\emptyset\{\frac{1}{\pi(X)}\mathrm{E}_\emptyset(TY \mid X)\} \\
&= \mathrm{E}_\emptyset\{\frac{1}{\pi(X)}\mathrm{E}_\emptyset(Y \mid X, T = 1)\mathrm{P}_\emptyset(T = 1 \mid X)\} \\
&= \mathrm{E}_\emptyset\{\mathrm{E}_\emptyset(Y \mid X, T = 1)\} = \mu_1 \qquad \text{by (36)}.
\end{aligned}$$

It automatically follows that $\mathrm{E}_\emptyset\{\frac{1-T}{1-\pi(X)}Y\} = \mu_0$. By Lemma 4, we see that, under the observational regime, $\frac{T}{\pi(X)}Y$ and $\frac{1-T}{1-\pi(X)}Y$ are unbiased estimators of μ_1 and μ_0 respectively.

One may have noticed that the two terms for ACE in (37) are similar with the Horvitz-Thompson (HT) estimator for sample surveys [12]. They are, however, different in various aspects. The aim of HT estimator is to estimate the mean of a finite population $Y_1, ..., Y_N$, denoted by $\mu = N^{-1}\sum_{i=1}^{N} Y_i$, from a stratified sample of size n drawn without replacement. For $i = 1, ..., N$, let Δ_i be binary sampling indicator ($\Delta_i = 1$: unit i is in sample; 0: unit i is not in sample), and π_i be the probability that unit i being drawn in the sample. Then HT estimator is given by:
$$\widehat{\mu}_{HT} = N^{-1}\sum_{i=1}^{N}\frac{\Delta_i}{\pi_i}Y_i, \tag{38}$$

where π_i is pre-specified, and thus known in a sample survey design. But the propensity model $\pi(X)$ in (37) is normally unknown. Moreover, HT estimator is applied to estimate the mean of a finite population, while $\widehat{\text{ACE}}$ is used to estimate the mean of a superpopulation [5]. HT estimator depends on pre-specified sampling scheme, but observations involved in $\widehat{\text{ACE}}$ are generated from, and thus are dependent on, the joint distribution of (X,T,Y) in the observational regime. Nevertheless, both HT estimator and $\widehat{\text{ACE}}$ are formed by means of the inverse probability weights $1/\pi_i$ or $1/\pi(X)$. In fact, HT estimator is also termed as the inverse probability weighted (IPW) estimator.

Sample surveys are closely related to missing data because the information is missing for those not sampled. So IPW estimator is frequently used in missing data models in the presence of partially observed response [1, 3, 14]. As counterfactuals are also regarded as missing data, IPW estimator can be used in the potential response framework with half observed information, to make causal inference of treatment effect under the assumptions of "strongly ignorable treatment assignment": $(Y(0),Y(1)) \perp\!\!\!\perp T \mid X$ and "no unobserved confounders" [1, 32].

5.1.1 Augmented inverse probability weighted estimator

From above discussion, there exists an unbiased estimator of ACE if either RRM or PM is correct. However, unknown RRM and PM makes it impossible to decide whether they are correct. Nevertheless, the augmented inverse probability weighted (AIPW) estimator can be constructed by combining the two models in the following alternative forms:

$$\begin{aligned}\widehat{\mu}_{1,\text{AIPW}} &= m(X) + \frac{T}{\pi(X)}(Y - m(X)) \\ &= \frac{T}{\pi(X)}Y + [1 - \frac{T}{\pi(X)}]m(X),\end{aligned} \tag{39}$$

and similarly,

$$\begin{aligned}\widehat{\mu}_{0,\text{AIPW}} &= m(X) + \frac{1-T}{1-\pi(X)}(Y - m(X)) \\ &= \frac{1-T}{1-\pi(X)}Y + [1 - \frac{1-T}{1-\pi(X)}]m(X),\end{aligned} \tag{40}$$

where $m(\cdot)$ and $\pi(\cdot)$ are arbitrary functions of X. As also indicated in its name, $\widehat{\mu}_{t,\text{AIPW}}$ is the sum of the IPW estimator and an augmented term.

Lemma 5. *Suppose that X is a strongly sufficient covariate. The estimator $\widehat{\mu}_{t,\text{AIPW}}$ has the property of **double robustness**. That is, $\widehat{\mu}_{t,\text{AIPW}}$ is an unbiased estimator of*

[5] In causal system, finite number of individuals in a study is called "population", which can be regardes as a sample from a larger "superpopulation" of interest

the population mean given $T = t$ by intervention, if either $\pi(X) = p_\emptyset(T = 1 \mid X)$ or $m(X) = E_\emptyset(Y \mid X, T = t)$.

Proof. By similarity, we only give proof of $\widehat{\mu}_{1,\mathrm{AIPW}}$. Consider the following two scenarios:

Scenario 1: $\pi(X) = p_\emptyset(T = 1 \mid X)$ and $m(X)$ is an arbitrary function of X.

It is easily seen that $\widehat{\mu}_{1,\mathrm{AIPW}}$ is unbiased, from the proof of Lemma 4. Since conditional on X, the last term in (39) vanishes when we take expectation of $\widehat{\mu}_{1,\mathrm{AIPW}}$ in the observational regime.

Scenario 2: $m(X) = \mathrm{E}_\emptyset(Y \mid X, T = 1)$ and $\pi(X)$ is an arbitrary function of X.

By (39), we have that

$$\begin{aligned}
\mathrm{E}(\widehat{\mu}_{1,\mathrm{AIPW}}) &= \mathrm{E}[m(X) + \frac{T}{\pi(X)}(Y - m(X))] \\
&= \mathrm{E}\{\mathrm{E}[m(X) \mid X]\} + \mathrm{E}\{\mathrm{E}[\frac{T}{\pi(X)}(Y - m(X)) \mid X]\} \\
&= \mathrm{E}[m(X)] + \mathrm{E}\{\frac{\mathrm{E}(TY \mid X) - m(X)\mathrm{E}(T \mid X)}{\pi(X)}\} \\
&= \mathrm{E}[m(X)] = \mu_1 \qquad \text{by (36).}
\end{aligned}$$

Indeed, if either $\pi(X) = p_\emptyset(T = 1 \mid X)$ or $m(X) = \mathrm{E}_\emptyset(Y \mid X, T = 1)$, not necessarily both, $\widehat{\mu}_{1,\mathrm{AIPW}}$ is unbiased. Consequently,

$$\widehat{\mathrm{ACE}}_{\mathrm{AIPW}} = \widehat{\mu}_{1,\mathrm{AIPW}} - \widehat{\mu}_{0,\mathrm{AIPW}}.$$

Theorem 4 *Suppose that X is a strongly sufficient covariate. Then the AIPW estimator $\widehat{\mathrm{ACE}}_{\mathrm{AIPW}}$ is doubly robust.*

To prove Theorem 4, we simply apply the fact that both $\widehat{\mu}_{1,\mathrm{AIPW}}$ and $\widehat{\mu}_{0,\mathrm{AIPW}}$ are doubly robust, so is their difference.

5.2 Parametric models

Suppose that we specify two parametric working models: the propensity working model $\pi(X;\alpha)$ and the response regression working model $m(T,X;\beta)$. Then by (39) and (40), we have, for the estimated $\mathrm{E}_1(Y)$ and $\mathrm{E}_0(Y)$, that

$$\widehat{\mu}_{1,\mathrm{AIPW}} = n^{-1}\{\sum_{i=1}^{n} \frac{T_i}{\pi(X_i;\hat{\alpha})}Y_i + [1 - \frac{T_i}{\pi(X_i;\hat{\alpha})}]m(1,X_i;\hat{\beta})\} \tag{41}$$

and

$$\widehat{\mu}_{0,\mathrm{AIPW}} = n^{-1}\{\sum_{i=1}^{n} \frac{1-T_i}{1-\pi(X_i;\hat{\alpha})} Y_i + [1 - \frac{1-T_i}{1-\pi(X_i;\hat{\alpha})}] m(0,X_i;\hat{\beta})\} \tag{42}$$

respectively. Therefore, by (41) and (42), we have that

$$\begin{aligned} \widehat{\mathrm{ACE}}_{AIPW} &= \hat{\mu}_{1,AIPW} - \hat{\mu}_{0,AIPW} \\ &= n^{-1}\{\sum_{i=1}^{n} [\frac{T_i}{\pi(X_i;\hat{\alpha})} - \frac{1-T_i}{1-\pi(X_i;\hat{\alpha})}](Y_i - m(T_i,X_i;\hat{\beta})\}, \end{aligned} \tag{43}$$

which is doubly robust, i.e., $\widehat{\mathrm{ACE}}_{AIPW}$ is a consistent and asymptotically normal estimator of ACE if either of the working models is correctly specified.

5.2.1 Discussion

Kang and Schafer [14] state that there are various ways to construct an estimator which is doubly robust. In our view, they are essentially the same, i.e., it must be in the same (or similar) form of AIPW estimator which is constructed by combining RRM and PM. Other constructions proposed in [14] are just variations of AIPW estimator. For example, in (38), instead of using N as denominator for each unit, they use normalised weights $\sum_{i=1}^{N} \frac{\Delta_i}{\pi_i}$. Such normalised weights are especially useful for precision improvement in the case that subjects with very small probabilities of being sampled are actually drawn from the population. Because if N is used as the weight, these subjects will influence the estimated average response enormously, and consequently, result in poor precision.

Kang and Schafer [14] have also investigated the precision performance of an doubly robust estimator when both $\pi(X)$ and $m(X)$ are moderately misspecified. They state that "in at least some settings, two wrong models are not better than one". This seems obvious because the performance of this estimator will depend on the degree of misspecification of both models. This can be easily analysed in theory but far more complicated in practice, as one can not have a good control of specifying models $\pi(X)$ and $m(X)$ based on limited observed data and previous experience (if any). Therefore, it would be difficult to measure to what extent the specified models are different from the true ones.

5.3 Precision of $\widehat{\mathrm{ACE}}_{AIPW}$

5.3.1 Known propensity score model

We already see that $\widehat{\mathrm{ACE}}_{AIPW}$ is an unbiased and doubly robust estimator of ACE. Then how can we choose an arbitrary function $m(X_i)$ to minimise the variance of $\widehat{\mathrm{ACE}}_{AIPW}$ given correct PM? Suppose that in an experiment, we know $\pi(X_i) = \mathrm{P}(T_i = 1 \mid X_i)$. Then in terms of the variance, we have that

$$\begin{aligned}
\mathrm{Var}(\widehat{\mathrm{ACE}}_{AIPW}) &= \mathrm{Var}\{n^{-1}[\sum_{i=1}^{n}(\frac{T_i}{\pi(X_i)}-\frac{1-T_i}{1-\pi(X_i)})(Y_i-m(X_i)]\} \\
&= n^{-2}\{\mathrm{Var}[\sum_{i=1}^{n}(\frac{T_i}{\pi(X_i)}-\frac{1-T_i}{1-\pi(X_i)})Y_i] \\
&\quad + \mathrm{Var}[\sum_{i=1}^{n}(\frac{T_i}{\pi(X_i)}-\frac{1-T_i}{1-\pi(X_i)})m(X_i)] \\
&\quad - 2\mathrm{Cov}[\sum_{i=1}^{n}(\frac{T_i}{\pi(X_i)}-\frac{1-T_i}{1-\pi(X_i)})Y_i, \sum_{i=1}^{n}(\frac{T_i}{\pi(X_i)}-\frac{1-T_i}{1-\pi(X_i)})m(X_i)]\} \\
&= n^{-2}\{\mathrm{Var}(\widehat{\mathrm{ACE}}_{HT})+\mathrm{E}[\sum_{i=1}^{n}\frac{m^2(X_i)}{\pi(X_i)(1-\pi(X_i))}] \\
&\quad - 2\mathrm{E}[\sum_{i=1}^{n}\frac{m(X_i)\mu_{1i}}{\pi(X_i)(1-\pi(X_i))}-\frac{m(X_i)(\mu_{1i}-\mu_i)}{(1-\pi(X_i))^2}] \\
&= n^{-2}\{\mathrm{Var}(\widehat{\mathrm{ACE}}_{HT})+\mathrm{E}[\sum_{i=1}^{n}\frac{m^2(X_i)}{\pi(X_i)(1-\pi(X_i))} \\
&\quad - 2\sum_{i=1}^{n}\{\frac{\mu_{1i}}{\pi(X_i)(1-\pi(X_i))}-\frac{\mu_{1i}-\mu_i}{(1-\pi(X_i))^2}\}m(X_i)]\},
\end{aligned}$$

where $\mu_{1i}=\mathrm{E}_\emptyset(Y_i \mid X_i, T_i=1)$ and $\mu_i=\mathrm{E}_\emptyset(Y_i \mid X_i)$.

By minimising the quadratic function of $m(X_i)$ in the expectation, it follows that

$$\begin{aligned}
m(X_i) &= [1-\pi(X_i)]\mu_{1i}+\pi(X_i)\mu_{0i} \\
&= [1-\pi(X_i)]\mathrm{E}_\emptyset(Y_i \mid X_i, T_i=1)+\pi(X_i)\mathrm{E}_\emptyset(Y_i \mid X_i, T_i=0), \qquad (44)
\end{aligned}$$

which minimises the variance of $\widehat{\mathrm{ACE}}_{AIPW}$ among all functions of X_i. In fact, if either $\pi(X_i)=p_\emptyset(T_i=1 \mid X_i)$, or (44) holds, $\widehat{\mathrm{ACE}}_{AIPW}$ is unbiased, and thus is doubly robust.

Let $m_1(X_i)$ and $m_0(X_i)$ denote the regressions of Y on X_i for the two treatment groups in the observational regime. It is unnecessary to require that $m_1(X_i)=\mathrm{E}_\emptyset(Y_i \mid X_i, T_i=1)$ and that $m_0(X_i)=\mathrm{E}_\emptyset(Y_i \mid X_i, T_i=0)$. As long as $m(X_i)$ is specified as the sum of the weighted expectations as in the form of (44), $m(X_i)$ minimises the variance of the estimated ACE.

Same result is obtained in [29] as (44), by minimising a weighted mean squared error of $m(X_i)$. We now discuss an alternative approach provided in [29]. Let $\tilde{Y}_i$ denote a weighted response in a form as follows:

$$\tilde{Y}_i=[\{\frac{1}{\pi(X_i)}-1\}T_i+\{\frac{1}{1-\pi(X_i)}-1\}(1-T_i)]Y_i. \qquad (45)$$

Then by (44), it follows that

$$m(X_i) = \frac{1-\pi(X_i)}{\pi(X_i)}\mathrm{E}_\emptyset(Y_i \mid X_i, T_i=1)\mathrm{P}(T=1 \mid X)$$

$$
\begin{aligned}
&+ \frac{\pi(X_i)}{1-\pi(X_i)}\mathrm{E}_\emptyset(Y_i \mid X_i, T_i = 0)\mathrm{P}(T = 0 \mid X) \\
&= \frac{1-\pi(X_i)}{\pi(X_i)}\mathrm{E}_\emptyset(T_i Y_i \mid X_i) + \frac{\pi(X_i)}{1-\pi(X_i)}\mathrm{E}_\emptyset[(1-T_i)Y_i \mid X_i] \\
&= \mathrm{E}_\emptyset\{[\frac{1-\pi(X_i)}{\pi(X_i)}T_i + \frac{\pi(X_i)}{1-\pi(X_i)}(1-T_i)]Y_i \mid X_i\} \\
&= \mathrm{E}_\emptyset\{[(\frac{1}{\pi(X_i)}-1)T_i + (\frac{1}{1-\pi(X_i)}-1)(1-T_i)]Y_i \mid X_i\} \\
&= \mathrm{E}_\emptyset(\tilde{Y}_i \mid X_i),
\end{aligned}
$$

where $m(X_i)$ is obtained by simply regressing $\tilde{Y}_i$ on X_i, rather than regressing Y_i on both X_i and T_i. However, an obvious disadvantage of this approach is its low precision. When individuals with the PS close to 0 are actually in the treatment group and/or those with the PS close to 1 are actually assigned to the control group, the weights $1/\pi(X_i)$ or $1/(1-\pi(X_i))$ of these units will be very large, which leads to corresponding responses being highly influential, which is dangerous. In fact, it may be even worse than the HT estimator as we will see next.

To show the difference of these approaches, we have implemented Monte Carlo computations for four estimators of $\widehat{\mathrm{ACE}}_{AIPW}$:

1. by (44) with $\mathrm{E}_\emptyset(Y_i \mid X_i, T_i = 1)$ and $\mathrm{E}_\emptyset(Y_i \mid X_i, T_i = 0)$ estimated by regressing Y_i on (X_i, T_i).
2. by (44) with $\mathrm{E}_\emptyset(Y_i \mid X_i, T_i = 1)$ and $\mathrm{E}_\emptyset(Y_i \mid X_i, T_i = 0)$ estimated by regressing Y_i on X_i for the treatment group and control group separately.
3. by Horvitz-Thompson approach, i.e. without covariate adjustment.
4. by regression of $\tilde{Y}_i$ on X_i as in (46).

The results of simulated 100 datasets are shown in Fig. 10. The first two approaches give similar results. That is, we can estimate $\mathrm{E}_\emptyset(Y_i \mid X_i, T_i = 1)$ and $\mathrm{E}_\emptyset(Y_i \mid X_i, T_i = 0)$ either simultaneously from the response regression on the treatment and X, or separately from the response regression only on X for each treatment group. As expected, the last approach generates several extreme estimates relative to others, which makes its variance even much larger than that of the HT estimator.

5.3.2 Known response regression model

Suppose that $\mathrm{E}_\emptyset(Y_i \mid X_i, T_i = 1)$ and $\mathrm{E}_\emptyset(Y_i \mid X_i, T_i = 0)$ are both known but not the PM. Then the AIPW estimator can be constructed as:

$$
\widehat{\mathrm{ACE}}_{AIPW} = n^{-1}\{\sum_{i=1}^{n}[\frac{T_i}{g(X_i)} - \frac{1-T_i}{1-g(X_i)}](Y_i - m(X_i))\},
$$

where

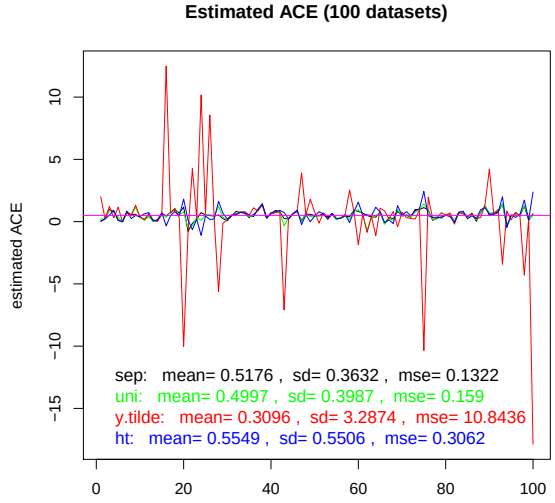


Fig. 10 Precision of the estimated ACE based on: (1) specified model for $\mathrm{E}_{\emptyset}(Y_i \mid X_i, T_i)$; (2) specified models for $\mathrm{E}_{\emptyset}(Y_i \mid X_i)$ separately for both groups; (3) Horvitz-Thompson estimator; (4) regression of $\tilde{Y}_i$ on X_i.

$$m(X_i) = (1-g(X_i))\mathrm{E}(Y_i \mid X_i, T_i = 1) + g(X_i)\mathrm{E}(Y_i \mid X_i, T_i = 0),$$

and $g(X_i)$ is an arbitrary function of X_i.

So $\widehat{\mathrm{ACE}}_{AIPW}$ is unbiased and its variance is computed as follows.

$$\begin{aligned}
\mathrm{Var}(\widehat{\mathrm{ACE}}_{AIPW}) &= \mathrm{Var}\{n^{-1}[\sum_{i=1}^{n}(\frac{T_i}{g(X_i)} - \frac{1-T_i}{1-g(X_i)})(Y_i - m(X_i)]\} \\
&= n^{-2}\mathrm{Var}\{\sum_{i=1}^{n}(\frac{T_i}{g(X_i)} - \frac{1-T_i}{1-g(X_i)})(Y_i - [1-g(X_i))\mu_{1i} + g(X_i)\mu_{0i}]\} \\
&= n^{-2}\mathrm{Var}\{\sum_{i=1}^{n}(\mu_{1i} - \mu_{0i}) + \frac{T_i}{g(X_i)}(Y_i - \mu_{1i}) - \frac{1-T_i}{1-g(X_i)}(Y_i - \mu_{0i})\} \\
&= n^{-2}\mathrm{Var}\{\sum_{i=1}^{n}(\mu_{1i} - \mu_{0i})\} \\
&\quad + n^{-2}\mathrm{E}\{\mathrm{Var}[\sum_{i=1}^{n}\frac{T_i}{g(X_i)}(Y_i - \mu_{1i}) - \frac{1-T_i}{1-g(X_i)}(Y_i - \mu_{0i}) \mid X_i]\} \\
&> n^{-2}\mathrm{Var}\{\sum_{i=1}^{n}(\mu_{1i} - \mu_{0i})\} = \mathrm{Var}(\widehat{\mathrm{ACE}}_{RRM}).
\end{aligned}$$

Hence, we conclude that, for each individual, if the conditional expectations of the response given X_i for both groups are known or correctly specified, then $\widehat{\text{ACE}}_{AIPW}$ will be less precise than the estimated ACE from the response regressions.

5.3.3 Discussion

If the PM is known, then the variance of $\widehat{\text{ACE}}_{AIPW}$ is minimised when $m(X_i)$ is specified as in (44) – where separate specification of $m_1(X_i)$ and $m_0(X_i)$ are not necessary. Rubin and van de Laan [29] has introduced a weighted response serving as an alternative, but we have shown, by simulations, that it could result in large variance of the estimated ACE and possibly larger than the HT estimator. In the case that the RRM is correctly specified, i.e., $m_1(X_i) = \text{E}_\emptyset(Y_i \mid X_i, T_i = 1)$ and $m_0(X_i) = \text{E}_\emptyset(Y_i \mid X_i, T_i = 0)$, then these two models rather than the AIPW estimator should be used to estimate ACE for higher precision of the estimator.

6 Summary

In this chapter, we have addressed statistical causal inference using Dawid's decision-theoretic framework within which assumptions are, in principle, testable. Throughout, the concept of sufficient covariate plays a crucial role. We have investigated propensity analysis in a simple normal linear model, as well as in logistic model, theoretically and by simulation. Adding weight to previous evidence [10, 21, 11, 33, 31], our results show that propensity analysis does little in improving estimation of the treatment causal effect, either unbiasedness or precision. However, as part of the augmented inverse probability weighted estimator that is doubly robust, correct propensity score model helps provide unbiased average causal effect.

Appendix

R code of simulations and data analysis

```
######################################################################
Figure 5: Linear regression (homoscedasticity)
----------------------------------------------------------------------
1. Y on X;
2. Y on population linear discriminant / propensity variable LD;
3. Y on sample linear discriminant / propensity variable LD*;
4. Y on population linear predictor LP.
######################################################################
```

```
##  set parameters

p <- 2
delta <- 0.5
phi <- 1
n <- 20

alpha <- matrix(c(1,0), nrow=1)
sigma <- diag(1, nrow=p)
b <- matrix(c(0,1), nrow=p)


##  create a function to compute ACE from four linear regressions

ps <- function(r) {

  #  data for T, X and Y from the specified linear normal model

  set.seed(r)
  .Random.seed
  t <- rbinom(n, 1, 0.5)

  require(MASS)
  m <- rep(0, p)
  ex <- mvrnorm(n, mu=m, Sigma=sigma)
  x <- t%*%alpha + ex

  ey <- rnorm(n, mean=0, sd=sqrt(phi))
  y <- t*delta + x%*%b + ey

  #  calculate the true and sample linear discriminants

  ld.true <- x%*%solve(sigma)%*%t(alpha)
  pred <- x%*%b

  d1 <- data.frame(x, t)
  c <- coef(lda(t~.,d1))
  ld <- x%*%c

  #  extract estimated average causal effect (ACE)
  #  from the four linear regressions

  dhat.pred <- coef(summary(lm(y~1+t+pred)))[2]
  dhat.x <- coef(summary(lm(y~t+x)))[2]
```

```
  dhat.ld <- coef(summary(lm(y~t+ld)))[2]
  dhat.ld.true <- coef(summary(lm(y~t+ld.true)))[2]

  return(c(dhat.x, dhat.ld, dhat.ld.true, dhat.pred))
}


##  estimate ACE from 200 simulated datasets
##  compute mean, standard deviation and mean square error of ACE

g <- rep(0, 4)
for (r in 31:230) {
  g <- rbind(g, ps(r))
}
g <- g[-1,]

d.mean <- 0
d.sd <- 0
mse <- 0

for (i in 1:4) {
  d.mean[i] <- round(mean(g[,i]),4)
  d.sd[i] <- round(sd(g[,i]),4)
  mse[i] <- round((d.sd[i])^2+(d.mean[i]-delta)^2, 4)
}


##  generate Figure 5

par(mfcol=c(2,2), oma=c(1.5,0,1.5,0), las=1)
main=c("M0:  Y on (T, X=(X1, X2)')", "M3:  Y on (T, LD*)",
    "M1: Y on (T, LD=X1)", "M2:  Y on (T, LP=X2)")

for (i in 1:4){
  hist(g[,i], br=seq(-2.5, 2.5, 0.5), xlim=c(-2.5, 2.5), ylim=c(0,80),
       main=main[i], col.lab="blue", xlab="", ylab="",col="magenta")
  legend(-2.5,85, c(paste("mean = ",d.mean[i]), paste("sd = ",d.sd[i]),
       paste("mse = ",mse[i])), cex=0.85, bty="n")
}
mtext(side=3, cex=1.2, line=-1.1, outer=T, col="blue",
    text="Linear regression (homoscedasticity) [200 datasets]")

dev.copy(postscript,"lrpvpdecmbook.ps", horiz=TRUE, paper="a4")
dev.off()
```

```
##############################################################################
Linear regression and subclassification (heteroscedasticity)
------------------------------------------------------------------------------

Figure 6:
1. Regression on population linear predictor LP;
2. Regression on population linear discriminant LD;
3. Regression on population quadratic discriminant / propensity variable QD;
4. Subclassification on QD.

Figure 7:
1. Regression on sample linear predictor LP*;
2. Regression on sample linear discriminant LD*;
3. Regression on sample quadratic discriminant / propensity variable QD*;
4. Subclassification on QD*.
##############################################################################

##  set parameters

p <- 20
d <- 0
delta <- 0.5
phi <- 1
n <- 500

a <- matrix(rep(0,p), nrow=1)
alpha <- matrix(c(0.5,rep(0,p-1)), nrow=1)
sigma1 <- diag(1, nrow=p)
sigma0 <- diag(c(rep(0.8, 10), rep(1.3, 10)), nrow=p)
b <- matrix(c(0, 1, rep(0,p-2)), nrow=p)


##  create a function to compute ACE from eight approaches

ps <- function(r) {

  #  data for T, X and Y from the specified linear normal model

  set.seed(r)
  .Random.seed
  pi <- 0.5
  t <- rbinom(n, 1, pi)
  n0 <- 0
```

```
for (i in 1:n) {
  if (t[i]==0)
  n0 <- n0+1
}

t <- sort(t, decreasing=FALSE)
mu1 <- a+alpha
mu0 <- a

require(MASS)
m <- rep(0, p)
ex0 <- mvrnorm(n0, mu=m, Sigma=sigma0)
ex1 <- mvrnorm((n-n0), mu=m, Sigma=sigma1)

a <- matrix(rep(a, n), nrow=n, byrow=TRUE)
x0 <- a[(1:n0),] + t[1:n0]%*%alpha + ex0
x1 <- a[(n0+1):n,] + t[(n0+1):n]%*%alpha + ex1
x <- rbind(x0, x1)

ey <- rnorm(n, mean=0, sd=sqrt(phi))
d <- rep(d, n)
y <- d + t*delta + x%*%b + ey

#  calculate linear discrimant, quadratic discrimant, for population
#  and for sample, extract estimated ACE from linear regressions

ld <- x%*%solve(pi*sigma1+pi*sigma0)%*%t(alpha)
d1 <- data.frame(x, t)
c <- coef(lda(t~.,d1))
ld.s <- x%*%c

z1 <- x%*%(solve(sigma1)%*%t(mu1) - solve(sigma0)%*%t(mu0))
z2 <- 0
for (j in 1:n){
  z2[j] <-  - 1/2*matrix(x[j,], nrow=1)%*%(solve(sigma1)
    - solve(sigma0))%*%t(matrix(x[j,], nrow=1))
}
qd <- z1+z2

dhat.x2 <- coef(summary(lm(y~1+t+x[,2])))[2]
dhat.ld <- coef(summary(lm(y~1+t+ld)))[2]
dhat.qd <- coef(summary(lm(y~1+t+qd)))[2]

mn <- aggregate(d1, list(t=t), FUN=mean)
m0 <- as.matrix(mn[1, 2:(p+1)])
```

```
  m1 <- as.matrix(mn[2, 2:(p+1)])
  v0 <- var(x0)
  v1 <- var(x1)

  c1 <- solve(v1)%*%t(m1)-solve(v0)%*%t(m0)
  z1.s <- x%*%c1
  c2 <- solve(v1)-solve(v0)
  z2.s <- 0
  for (i in 1:n){
    z2.s[i] <- -1/2*matrix(x[i,], nrow=1)%*%c2%*%t(matrix(x[i,], nrow=1))
  }
  qd.s <- z1.s+z2.s

  dhat.x <- coef(summary(lm(y~1+t+x)))[2]
  dhat.ld.s <- coef(summary(lm(y~1+t+ld.s)))[2]
  dhat.qd.s <- coef(summary(lm(y~1+t+qd.s)))[2]

  #  extract estimated ACE from subclassification

  d2 <- data.frame(cbind(qd, qd.s, y, t))

  tm1 <- vector("list", 2)
  tm0 <- vector("list", 2)
  te.qd <- 0

  for (k in 1:2) {
    d3 <- d2[, c(k,3,4)]
    d3 <- split(d3[order(d3[,1]), ], rep(1:5, each=100))

    tm <- vector("list", 5)
    for (j in 1:5) {
      tm[[j]] <-   aggregate(d3[[j]], list(Stratum=d3[[j]]$t), FUN=mean)
      tm1[[k]][j] <- tm[[j]][2,3]
      tm0[[k]][j] <- tm[[j]][1,3]
    }
    te.qd[k] <- sum(tm1[[k]] - tm0[[k]])/5
  }

#  return estimated ACE from the eight approaches

  return(c(dhat.x2, te.qd[1], dhat.ld, dhat.qd,
    dhat.x, te.qd[2], dhat.ld.s, dhat.qd.s))
}
```

```
##  estimate ACE from 200 simulated datasets
##  compute mean, standard deviation and mean square error of ACE

g <- rep(0, 8)
for (r in 31:230) {
   g <- rbind(g, ps(r))
}
g <- g[-1,]

d.mean <- 0
d.sd <- 0
d.mse <- 0

for (i in 1:8) {
  d.mean[i] <- round(mean(g[,i]),4)
  d.sd[i] <- round(sd(g[,i]),4)
  d.mse[i] <- round((d.sd[i])^2+(d.mean[i]-delta)^2, 4)
}


##  generate Figure 6

par(mfcol=c(2,2), oma=c(1.5,0,1.5,0), las=1)
main=c("Regression on LP=X2","Subclassification on QD",
    "Regression on  LD=5/9X1","Regression on QD")
for (i in 1:4){
  hist(g[,i], br=seq(-0.1, 1.1, 0.1), xlim=c(-0.1, 1.1), ylim=c(0,80),
      main=main[i], col.lab="blue", xlab="", , ylab="", col="magenta")
  legend(-0.2,85, c(paste("mean = ",d.mean[i]), paste("sd = ",d.sd[i]),
      paste("mse = ",d.mse[i])), cex=0.85, bty="n")
}
mtext(side=3, cex=1.2, line=-1.1, outer=T, col="blue",
    text="Linear regression and subclassification
    (heteroscedasticity) [200 datasets]")

dev.copy(postscript,"pslrsubtruebook.ps", horiz=TRUE, paper="a4")
dev.off()


##  generate Figure 7
main=c("Regression on X","Subclassification on QD*",
    "Regression on LD*",  "Regression on QD*")
for (i in 1:4){
  hist(g[,i+4], br=seq(-0.1, 1.1, 0.1), xlim=c(-0.1,1.1), ylim=c(0,80),
      main=main[i], col.lab="blue", xlab="", ylab="", col="magenta")
```

```
  legend(-0.2,85, c(paste("mean = ",d.mean[i+4]), paste("sd = ",d.sd[i+4]),
      paste("mse = ",d.mse[i+4])), cex=0.85, bty="n")
}
mtext(side=3, cex=1.2, line=-1.1, outer=T, col="blue",
    text="Linear regression and subclassification
    (heteroscedasticity, sample) [200 datasets]")

dev.copy(postscript,"pslrsubbook.ps", horiz=TRUE, paper="a4")
dev.off()



#########################################################################
Figure 9 and Table 1: Propensity analysis of custodial sanctions study
-------------------------------------------------------------------------
1. Y on X;
2. Y on population linear discriminant / propensity variable LD;
3. Y on sample linear discriminant / propensity variable LD*;
4. Y on population linear predictor LP.
#########################################################################

##  read data, imputation by bootstrapping for missing data

dAll = read.csv(file="pre_impute_data.csv", as.is=T, sep=',', header=T)

set.seed(100)
.Random.seed
library(mi)
data.imp <- random.imp(dAll)


## estimate propensity score by logistic regression

glm.ps<-glm(Sentenced_to_prison~
                Age_at_1st_yuvenile_incarceration_y +
                N_prior_adult_convictions +
                Type_of_defense_counsel +
                Guilty_plea_with_negotiated_disposition +
                N_jail_sentences_gr_90days +
                N_juvenile_incarcerations +
                Monthly_income_level +
                Total_counts_convicted_for_current_sentence +
                Conviction_offense_type +
                Recent_release_from_incarceration_m +
                N_prior_adult_StateFederal_prison_terms +
```

```
                Offender_race +
                Offender_released_during_proceed +
                Separated_or_divorced_at_time_of_sentence +
                Living_situation_at_time_of_offence +
                Status_at_time_of_offense +
                Any_victims_female,
                data = data.imp, family=binomial)

summary(glm.ps)
eps <- predict(glm.ps, data = data.imp[, -1], type='response')
d.eps <- data.frame(data.imp, Est.ps = eps)


## Figure 9: densities of estimated propensity score (prison vs. probation)

library(ggplot2)

d.plot <- data.frame(Prison = as.factor(data.imp$Sentenced_to_prison),
    Est.ps = eps)
pdf("ps.dens.book.pdf")
ggplot(d.plot, aes(x=Est.ps, fill=Prison)) + geom_density(alpha=0.25) +
    scale_x_continuous(name="Estimated propensity score") +
    scale_y_continuous(name="Density")
dev.off()


## logistic regression of the outcome on all 17 variables

glm.y.allx<-glm(Recidivism~
                Sentenced_to_prison +
                Age_at_1st_yuvenile_incarceration_y +
                N_prior_adult_convictions +
                Type_of_defense_counsel +
                Guilty_plea_with_negotiated_disposition +
                N_jail_sentences_gr_90days +
                N_juvenile_incarcerations +
                Monthly_income_level +
                Total_counts_convicted_for_current_sentence +
                Conviction_offense_type +
                Recent_release_from_incarceration_m +
                N_prior_adult_StateFederal_prison_terms +
                Offender_race +
                Offender_released_during_proceed +
                Separated_or_divorced_at_time_of_sentence +
                Living_situation_at_time_of_offence +
```

```
                Status_at_time_of_offense +
                Any_victims_female,
                data = d.eps, family=binomial)

summary(glm.y.allx)


## logistic regression of the outcome on the estimated propensity score

glm.y.eps<-glm(Recidivism ~ Sentenced_to_prison + Est.ps,
  data = d.eps, family=binomial)
summary(glm.y.eps)
```

References

1. Bang H., Robins J.M.: Doubly robust estimation in missing data and causal inference models. Biometrics **61**, 962–972 (2005)
2. Berzuini G.: Causal inference methods for criminal justice data, and an application to the study of the criminogenic effect of custodial sanctions. MSc Thesis in Applied Statistics. Birkbeck College, University of London (2013)
3. Carpenter J.R., Kenward M.G., Vansteelandt S.: A comparison of multiple imputation and doubly robust estimation for analyses with missing data. J. R. Stat. Soc. Series A **169**, 571–584 (2006)
4. Dawid A.P.: Conditional independence in statistical theory (with discussion). J. R. Stat. Soc. Series B **41**, 1–31 (1979)
5. Dawid A.P.: Conditional independence for statistical operations. Ann. Stat. **8**, 598–617 (1980)
6. Dawid A.P.: Causal inference without counterfactuals. J. Amer. Statist. Ass. **95**, 407–424 (2000)
7. Dawid A.P.: Influence diagrams for causal modelling and inference. Int. Stat. Rev. **70**, 161–189 (2002)
8. Fisher R.A.: Theory of statistical estimation. Proc. Cam. Phil. Soc. **22**, 700–725 (1925)
9. Guo H., Dawid A.P.: Sufficient covariates and linear propensity analysis. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), Chia Laguna, Sardinia, Italy, May 13-15, 2010, edited by Y. W. Teh and D. M. Titterington. Journal of Machine Learning Research Workshop and Conference Proceedings **9**, 281–288 (2010)
10. Hahn J.: On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica **66**, 315–331 (1998)
11. Hirano K., Imbens G.W., Ridder G.: Efficient estimation of average treatment effects using the estimated propensity score. Econometrica **71**, 1161–1189 (2003)
12. Horvitz D.G., Thompson D.J.: A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Ass. **47**, 663–685 (1952)
13. Imbens G.W., Lemieux T.: Regression discontinuity designs: A guide to practice. J. Econometrics **142**, 615–635 (2007)
14. Kang J.D.Y., Schafer J.L.: Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Statist. Sci. **22**, 523–539 (2007)

15. Mardia K.V., Kent J.T., Bibby J.M.: Multivariate Analysis. Academic Press, New York (1979)
16. Pearl J.: Causal diagrams for empirical research (with Discussion). Biometrika **82**, 669–710 (1995)
17. Pearl J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2000)
18. Petersilia J.: Research in Brief: Probation and felony offenders. Washington, DC: National Institute of Justice (1985)
19. Petersilia J., Turner S., Kahan J.: Granting felons probation: Public risks and alternatives. Santa Monica: The Rand Corporation, R-3186-NIJ (1985)
20. Petersilia J, Turner S., Peterson J.: Prison versus probation in California: Implications for crime and offender recidivism. Santa Monica: The Rand Corporation, R-3323-NIJ (1986)
21. Robins J.M., Mark S.D., Newey W.K.: Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics **48**, 479–495 (1992)
22. Rosenbaum P.R., Rubin D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**, 44–55 (1983)
23. Rosenbaum P.R., Rubin D.B.: Reducing bias in observational studies using subclassification on the propensity score. J. Amer. Statist. Ass. **79**, 516–524 (1984)
24. Rubin D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**, 688–701 (1974)
25. Rubin D.B.: Assignment to treatment group on the basis of a covariate. J. Educ. Stat. **2**, 1–26 (1977)
26. Rubin D.B.: Bayesian inference for causal effects: the role of randomization. Ann. Stat. **6**, 34–68 (1978)
27. Rubin D.B.: Matched Sampling for Causal Effects. Cambridge University Press (2006)
28. Rubin D.B., Thomas N.: Characterizing the effect of matching using linear propensity score methods with normal distributions. Biometrika **79**, 797–809 (1992)
29. Rubin D.B., van de Laan M.J.: Covariate adjustment for the intention-to-treat parameter with empirical efficiency maximization. U.C.Berkeley Division of Biostatistics Working Paper 229 (2008)
30. Sekhon J.: Multivariate and propensity score matching software with automated balance optimization: The matching package for R. J. Statist. Software **42** (2011)
31. Senn S., Graf E., Caputo A.: Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. Stat. Med. **26**, 5529–5544 (2007)
32. Tang Z.: Understanding OR, PS, and DR, Comment on "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data" by Kang and Schafer. Statist. Sci. **22**, 560–568 (2007)
33. Winkelmayer W.C., Kurth T.: Propensity scores: Help or hype? Nephrol. Dial. Transplant. **19**, 1671–1673 (2004)